Université de Franche Comté – Besançon
Faculté des Sciences et Techniques
Département Informatique des Systèmes Complexes
Tuteur : Monsieur Bruno TATIBOUET

RÉALISATION D'UN DATAMART DÉCISIONNEL INTÉGRANT UN ETL OPEN SOURCE

Loubna HASSANI

Orange Labs Recherche et Développement
1 rue Maurice et Louis de Broglie
CS 20382
90007 BELFORT Cedex

Maître de stage :
Monsieur Thierry STUNAULT

Soutenu le 11 juin 2012

Remerciements

Un stage n'est jamais un travail individuel. Il s'inscrit dans une démarche globale de recherche qui bénéficie des travaux déjà réalisés, est en relation avec les travaux en cours et sera peut être repris par la suite.

Ce rapport doit donc énormément à M. Thierry STUNAULT qui m'a encadrée tout au long de cette période, car une bonne stagiaire n'est rien sans un bon maître de stage, ainsi que Mme Catherine CHEVANET, responsable de l'URD, grâce à qui cette opportunité de stage s'est présentée.

Je remercie mon encadrant de l'université M. Bruno TATIBOUET pour son encadrement constant et ses nombreux conseils.

Je remercie également M. Michel USCLADE et Mme Sandrine LOVISA pour leur part active dans cette initiative, et leurs conseils qui ont participés à l'avancement du projet.

Je tiens à remercier tout le personnel de l'équipe dont j'ai fait partie pendant cinq mois, à savoir l'Unité de Recherche et Développement ISA, « Information System Integration validation for Access network » d'Orange Labs Belfort.

Toutes ces personnes ont su rendre ce stage intéressant, d'abord au niveau professionnel, me permettant de découvrir des aspects techniques ou organisationnels que je ne connaissais pas, mais aussi humains, grâce un accueil chaleureux et une ambiance agréable.

Ce stage aura véritablement été un passage mémorable de ma formation, avant tout grâce à eux, et je les en remercie.

Table des matières

Table des illustrations

Introduction

Ce stage de fin d'études s'inscrit dans le cadre du Master Informatique professionnel Systèmes Distribués et Réseaux (SDR) de l'Université de Franche Comté. Il s'est déroulé au sein de l'unité de Recherche et Développement Orange Labs sur une période de cinq mois.

Le pilotage de l'entreprise est primordial dans le sens où il nécessite des choix qui consistent à dégager un profit durable. Il est important pour les performances de la société que ces prises de décisions soient basées sur l'état global de celle-ci. C'est ainsi qu'intervient le décisionnel, qui fournit une représentation intelligente des informations provenant des bases de données au travers d'outils spécialisés.

Dans un premier temps, je présenterai l'environnement professionnel dans lequel j'évolue. Je commencerai par une brève description des objectifs du groupe Orange, qui traduiront la nécessité d'une branche de Recherche & Développement efficace, innovante et proche des besoins du client. Je situerai ensuite l'unité dans laquelle je travaille par des descriptions successives de chaque niveau de l'organisation fonctionnelle de la branche Recherche & Développement.

Dans un deuxième temps je présenterais le projet « DataMart » qui m'a été confié, il est constitué de deux parties principales. La première partie consistait à faire une analyse poussée du DataMart existant. Des modifications de plusieurs natures étaient attendues, afin de rationaliser le traitement actuel (suppression des données inutiles et simplification des traitements…).

L'évolution de la chaîne de traitement poursuivait plusieurs objectifs : permettre un gain au niveau du temps d'exécution afin de fournir plus rapidement des informations cohérentes aux clients internes d'Orange, permettre éventuellement de réaliser une analyse plus fine de l'activité et enfin optimiser la base de données en ce qui concerne son contenu.

La seconde partie consistait, à partir de l'analyse effectuée dans la première partie, à réaliser un DataMart pour le suivi et l'analyse des données trafic ADSL des clients à partir des informations techniques ainsi que commerciales. Pour ce faire, l'outil open source Talend a été utilisé en remplacement de DataStage afin de le tester en grandeur nature. Le DataMart devait permettre d'obtenir une vision globale des activités des clients, et d'en assurer plus facilement le suivi et l'analyse.

Ce projet m'a permis en particulier d'être confronté aux problématiques de modélisation étudiées cette année. Il m'a offert également la possibilité de concevoir un DataMart dans sa totalité, de la phase d'étude à celle de l'alimentation en passant par l'analyse et la conception.

Chapitre 1 Cadre de stage

Dans le présent chapitre, j'introduis une présentation générale de l'entreprise où j'ai passé mon stage de fin d'étude.

Orange

1. Présentation

Orange est l'un des principaux opérateurs européens du mobile et de l'accès internet ADSL, et l'un des leaders mondiaux des services de télécommunications aux entreprises multinationales, sous la marque Orange Business Services.

Depuis 2006, Orange est la marque unique du groupe Orange pour l'Internet, la télévision et le mobile en France et dans la majorité des pays où le groupe est présent. Orange Business Services est la marque des services offerts aux entreprises dans le monde.

Orange est le 3ème opérateur mobile et 1er fournisseur d'accès Internet ADSL en Europe, il compte parmi les leaders mondiaux des services de télécommunications aux entreprises multinationales. Opérateur intégré, le groupe se donne les moyens d'être l'opérateur de référence des nouveaux services de télécommunications en Europe.

Chiffres clés d'Orange en France

- **100 000** salariés
- **7,4 millions** de Livebox
- **9 millions** de clients équipés haut débit, soit **46,3%** de part de marché Grand Public
- **26,2 millions** de clients mobiles dont **14,6** clients haut débit mobile
- **7,2 millions** de clients VOIP
- **1200** boutiques Orange
- **736 000** clients Orange TV
- **48 000 clients** ont signés pour la Fibre (644 000 foyers connectables)

Orange s'adresse simultanément à ses salariés, à ses clients, à ses actionnaires et plus largement à la société dans laquelle l'entreprise évolue en s'engageant concrètement

sur des plans d'actions. Ceux-ci concernent les salariés du groupe grâce à une nouvelle vision des Ressources Humaines ; les réseaux, avec le déploiement des infrastructures du futur sur lesquelles le groupe bâtira sa croissance ; les clients, avec l'ambition de leur offrir la meilleure expérience parmi les opérateurs grâce, notamment, à l'amélioration de la qualité de service ; et l'accélération du développement international.

Le groupe a changé aussi sa stratégie : il ne se contente plus uniquement de s'adapter aux technologies. La priorité identifiée aujourd'hui est de répondre au besoin des clients. Cette stratégie est à la fois une stratégie d'opérateur intégré, une stratégie de convergence et une stratégie d'innovation. Elle s'articule autour de trois modes d'actions :

- La simplicité, ou comment simplifier la vie des clients
- L'agilité, ou comment développer l'agilité du groupe dans l'exercice de ses métiers
- La performance durable, dont l'objectif est d'inscrire la performance dans la durée

2. Le grand projet : conquêtes 2015

Le comité de direction d'Orange a récemment changé. Ce changement, initialement prévu en 2012, a été précipité pour répondre à la crise sociale tant médiatisée dont le principal moteur est le stress vécu par ses salariés.

Le plus important changement concerne la nomination de Stéphane Richard au poste de Directeur Général. Le nouveau comité a décidé qu'il était nécessaire pour Orange de se remobiliser et d'avancer vers de nouveaux objectifs sociaux, financiers et organisationnels.

Cinq plans d'actions sont identifiés pour ce projet.

Les salariés au cœur du développement de l'entreprise : pour réaliser l'ambition d'Orange, elle doit réaliser ses engagements vis-à-vis de ses salariés, cela pour aboutir à une nouvelle vision des relations et des ressources humaines dans l'entreprise.

La montée en débit et la transformation des infrastructures du groupe : au bureau, à la maison ou en déplacement, les clients recherchent de plus en plus de contenus numériques et de communications interpersonnelles enrichies. Le groupe propose des moyens techniques pour assurer dans la durée l'acheminement de ces nouveaux services.

La simplicité et la fiabilité des produits et des services : c'est l'un des piliers de la qualité de service, avec la relation client. Simplicité et fiabilité constituent des leviers décisifs pour distinguer le groupe de ses concurrents et faire préférer Orange à l'ensemble des parties prenantes.

L'excellence dans la relation client : prospects, acheteurs, utilisateurs... Tous les clients doivent bénéficier d'une écoute et d'une relation irréprochables. L'action

d'Orange se fonde sur des principes simples, comme le respect des commandes et des délais de livraison par exemple. A l'entreprise d'enrichir la relation de moments privilégiés, adaptés au profil du client.

Les nouveaux services : le cœur de métier d'Orange est de créer du lien. Sa mission consiste à faciliter la vie numérique de ses clients. Pour cela, le groupe va leur proposer des nouvelles générations de services fiables et accessibles partout et quand ils le souhaitent.

Pays et entités ont adapté chaque plan d'action à la réalité de leurs marchés et de leurs métiers. Lors de ce travail de réflexion et de contribution, ils ont également fait émerger quatre principes communs à l'ensemble des salariés, des fondations sur lesquelles repose l'ensemble des transformations mises en œuvre.

L'engagement de tenir la promesse du groupe : montée en puissance des réseaux et des débits, accessibilité des offres, fiabilité et simplicité des services… Dans tous les plans d'action figure l'engagement des équipes à offrir le meilleur service au meilleur prix. Mais cela ne se traduit pas partout de la même façon : les métiers, les cultures et les équipes sont différents d'un pays et d'un marché à l'autre. Par conséquent, le groupe ne va pas formuler la même promesse à tous ses clients.

La volonté d'être un opérateur de confiance : Orange a lancé ou va lancer de nombreuses actions dans le domaine de la responsabilité sociale d'entreprise, la qualité de service de bout en bout, la connaissance de ses clients, la sécurité de leurs données et de leurs usages... Le sens de ces actions est tout simplement de développer la fidélité de ses clients – non seulement sur la base d'avantages tarifaires immédiats mais, dans la durée, en nouant une relation de confiance avec eux. Les clients sont le premier atout. A Orange de le valoriser, en facilitant et en protégeant la foisonnante vie numérique, en devenant leur opérateur de confiance.

L'innovation au bénéfice des clients : Orange est l'un des opérateurs de télécommunications les plus innovants au monde. Ses plans d'action et feuilles de route fourmillent d'initiatives pionnières pour mettre au point de nouvelles plates-formes de services, développer le « Cloud Computing », inventer de nouveaux usages dans les domaines de l'éducation, de la santé, de l'économie au quotidien… Ce qui intéresse le groupe, ce n'est pas la première marche du podium mondial des nouvelles technologies mais que, dans chaque pays où il est présent, ses clients le reconnaissent comme l'innovateur le plus actif dans leur environnement. Une entreprise qui, concrètement, améliore leur vie au quotidien.

Enfin, **être fier d'appartenir à Orange** : toutes les entités se sont mobilisées pour proposer des initiatives replaçant l'humain au cœur de la démarche du groupe. Ces initiatives sont multiples, comme le renouvellement des méthodes de management, la motivation et la responsabilisation des équipes, la reconnaissance des talents et de la performance, l'ouverture de nouvelles opportunités de carrière… Il s'en dégage un véritable élan commun, une envie de travailler ensemble, rassemblés avec fierté sous la bannière Orange. Le groupe a l'ambition de réunir toutes les conditions de la réussite pour chaque salarié(e) dans chacune de ses entités.

3. Orange Labs, la Recherche et Développement d'Orange

Orange Labs est, à l'image du groupe Orange, international. Il compte près de 3600 chercheurs et ingénieurs sur 17 sites répartis sur 4 continents.

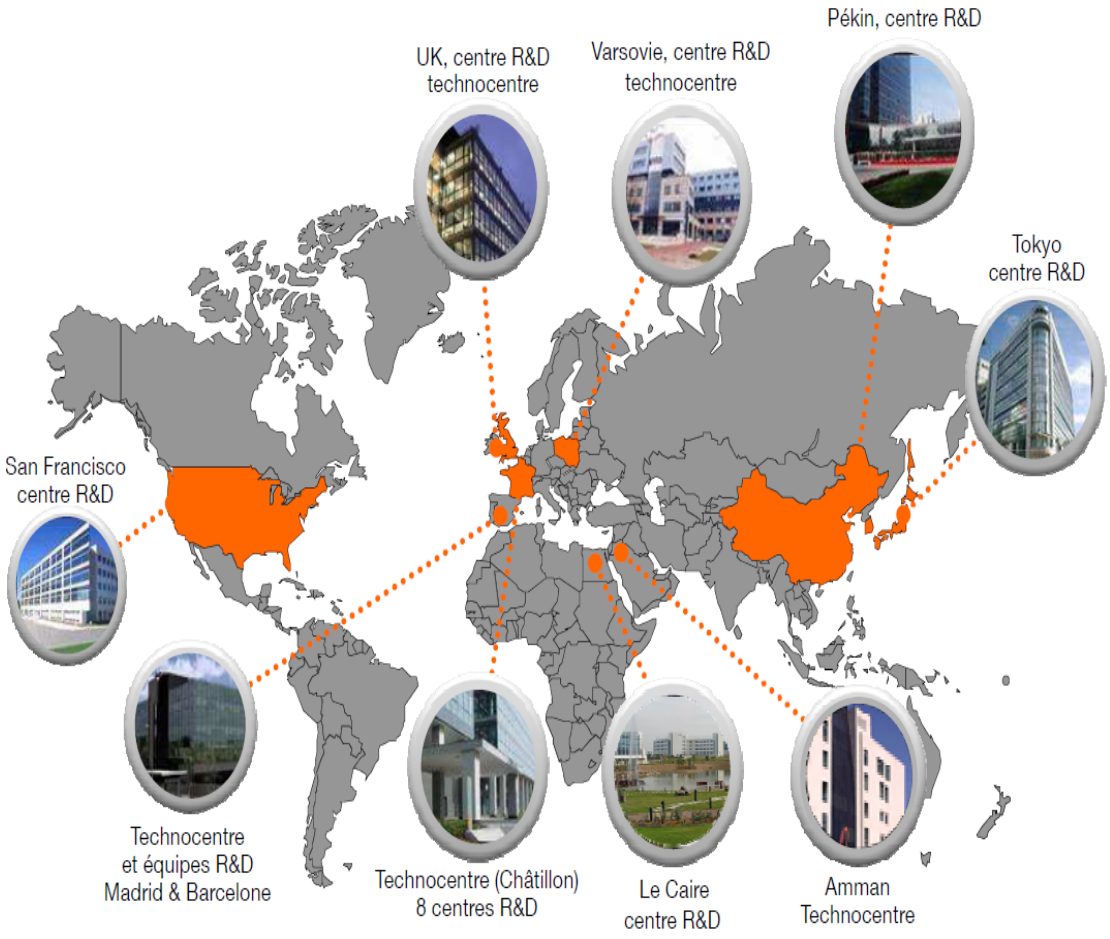


Figure 1 : LOCALISATION DES R&D A TRAVERS LE MONDE

La Division "Recherche et Développement" a pour principales missions :

- De développer des produits et services pour le groupe, en respectant la qualité de service
- De dégager de nouvelles sources de croissance
- D'anticiper les révolutions technologiques et d'usage
- D'imaginer dès maintenant les solutions du futur.

En contribuant à la convergence des technologies et à l'enrichissement des services d'Orange, la capacité de R&D constitue pour le Groupe un avantage stratégique majeur pour anticiper les grandes ruptures technologiques, orienter l'innovation du secteur des télécommunications et inventer la nouvelle génération de services : des services de communication intégrés, innovants et simples d'utilisation. Orange Labs présente deux dimensions, la recherche et le développement.

Ses activités de recherche permettent de détecter les ruptures technologiques, et acquérir un savoir-faire. Elles permettent également d'exceller en protection et valorisation de la propriété intellectuelle du groupe. Enfin, ce sont ces activités qui conduisent à l'exploration de nouvelles technologies, services et usages.

En termes de développement, Orange Labs doit concevoir les services du futur et améliorer les offres existantes, en réduisant les délais de mise sur le marché afin de répondre aux besoins du marché au plus tôt. Elle intervient également dans le développement de partenariats stratégiques avec les industriels. Elle représente le groupe dans les instances de normalisation des différentes normes technologiques.

La division Orange Labs est au cœur de la stratégie actuelle du groupe. A l'origine des produits proposés aux clients, elle doit mettre en avant la simplicité dans ces produits. Elle doit également s'approprier l'agilité nécessaire à l'évolution des technologies et des besoins. Dans le cadre de la performance durable, elle s'implique dans la mutualisation des réseaux, des plateformes et systèmes d'exploitation et dans la diffusion des innovations.

4. Organisation de la direction Recherche et Développement

La Division Recherche & Développement est composée en France, pour sa partie ingénierie technique, de six CRD (Centres de Recherche & Développement) structurés autour des types de services et de réseaux existants, comme le montre la Figure 3 :

- Le CRD Services Intégrés, Résidentiels et Personnels
- Le CRD Service aux Entreprises
- Le CRD Middleware et Plates-formes avancées
- Le CRD Réseaux d'accès
- Le CRD Technologies
- Le CRD Cœur de Réseau

Le CRD RÉSeaux d'Accès (RESA) est celui dans lequel j'effectue mon stage.

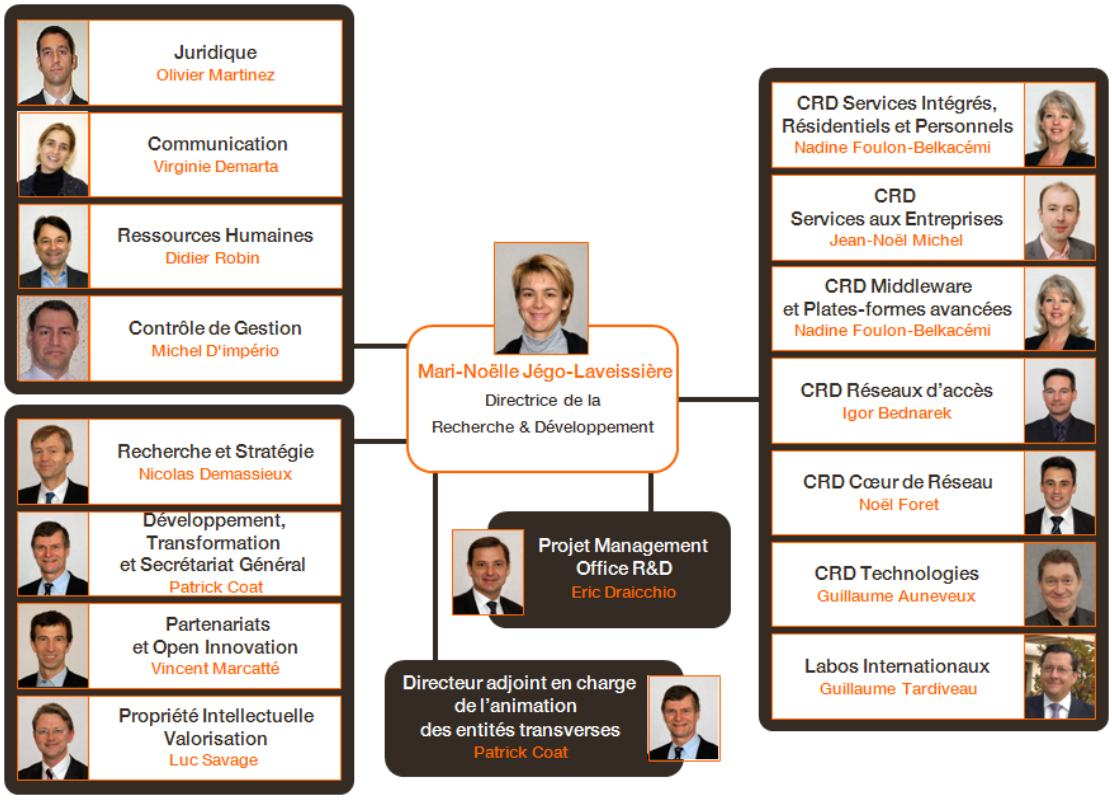


Figure 2 : ORGANIGRAMME ET STRUCTURE DE LA DIVISION R&D

5. Le CRD Réseaux d'Accès

Le CRD RESA réunit plus de 500 personnes (CDI + CDD) sur cinq sites différents.

Ce CRD a pour mission de :

- Développer les compétences clés pour intervenir efficacement sur les projets de développement et de recherche
- Identifier et construire les compétences de demain
- Conduire les projets et assurer la tenue des délais et la qualité des livrables

Il répartit, comme le présente la Figure 4, ces objectifs à travers quatre laboratoires :

- Fixed Access Network Architecture (ANA)
- Design and Evaluation of multiservices fixed Access Networks (DEAN)
- Performance & Engineering of the Access network (PEAK)
- Wireless Access Systems & Architecture (WASA)

Le laboratoire PEAK est celui dans lequel j'ai effectué mon stage.

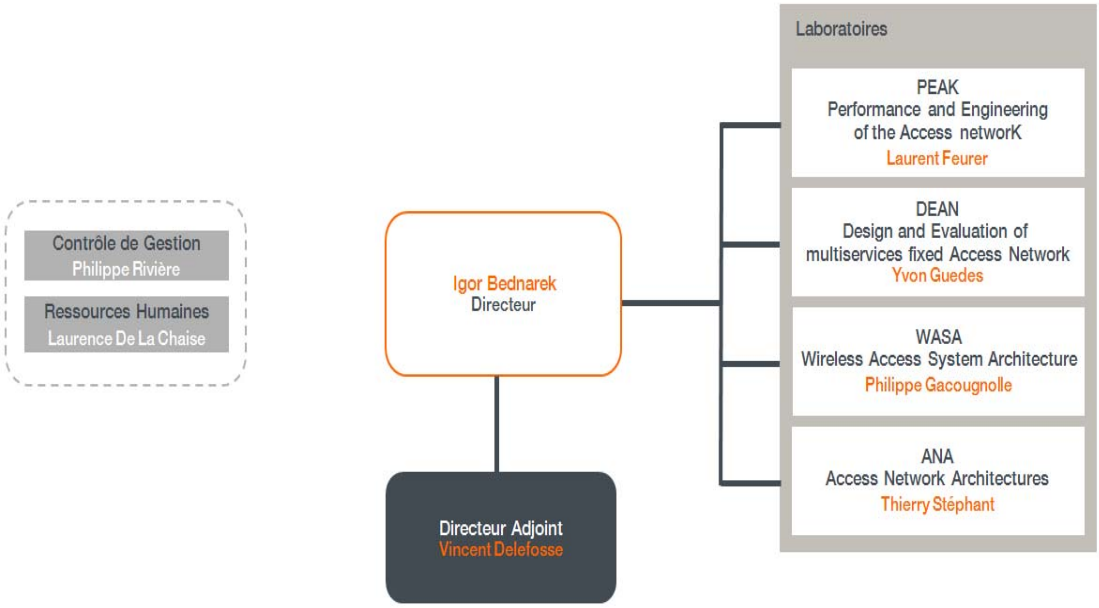


Figure 3 : ORGANISATION DU CRD RESA

6. Le laboratoire Performance & Engineering of the Access network (PEAK)

Le laboratoire PEAK a pour mission de définir des méthodes de planification et de développer des outils d'ingénierie du réseau d'accès mobile et fixe. Il intervient également sur les différentes phases du processus d'ingénierie : le dimensionnement, le déploiement et l'optimisation. Les méthodes et outils développés par le laboratoire PEAK permettent de caractériser le trafic, de dimensionner les équipements, d'améliorer la qualité de service et d'optimiser l'utilisation des ressources fixes et radio.

Enfin, ce laboratoire doit évaluer les coûts économiques de déploiement et d'exploitation du réseau pour aider au choix des meilleures solutions d'accès. Le laboratoire développe et diffuse ses outils dans le groupe Orange et les valorise auprès d'éditeurs de logiciels et de constructeurs d'infrastructure.

Le laboratoire PEAK est composé de cinq Unités de Recherche et Développement (URD) réparties sur les sites de Belfort et Issy-les-Moulineaux. C'est au sein de l'URD ISA (Information System, integration & validation for Access network), composée de 19 personnes et dirigée par Catherine CHEVANET que j'effectue mon stage.

7. L'unité Information System, integration & validation for Access network (ISA)

Cette URD effectue les études d'introduction des nouvelles technologies d'accès dans le Système d'Information (SI) d'Orange. D'autre part, elle contribue à la qualité des logiciels produits par le laboratoire PEAK en accompagnant le déploiement des processus qualité. Ces processus permettent de valider la qualité des outils développés.

Ses compétences sont diverses. En termes de production de logiciels, elle est qualifiée en développement, en validation et en mise en place de processus qualité. A cela s'ajoute sa connaissance du Système d'Information réseau, le datamining (capacité à extraire des informations à partir de données), la cartographie et la conception d'architectures techniques.

Figure 4 : ORGANISATION DE L'UNITE ISA

8. Organisation des projets

Les projets au sein d'Orange Labs sont organisés de façon hiérarchique et thématique. La cohérence, la performance et la mise en visibilité des activités de recherche et de développement sont garanties par deux entités différentes.

La première, la Direction Recherche et Stratégie, supervise le contenu, le budget et la répartition des projets côté recherche. La seconde, la Direction du Développement, occupe les mêmes responsabilités côté développement. Le budget des projets de recherche est géré en interne, alors que les budgets des projets de développements sont couverts par les financements des clients des projets développés.

La répartition des projets est faite au travers de programmes pour les projets de développement, et d'objets de recherche pour les projets de recherche.

Les programmes/objets de recherche regroupent des projets et des travaux courants selon leur domaine, par exemple les outils d'ingénierie réseau, et sont composés de tâches. Les travaux courants sont différents des projets, qui sont définis par un budget, un périmètre, un délai et éventuellement une qualité. Il n'y a pas d'engagement de livrable pour les travaux courants, qui spécifient uniquement un besoin de moyens pour une durée donnée. Les travaux courants couvrent typiquement les activités de maintenance et de support.

La gestion de ces différents éléments se fait à chaque niveau : il y a des responsables de programme et d'objet de recherche, des chefs de projets de développement et de recherche, des responsables de travaux courants et des responsables de tâches.

Chapitre 2 Descriptif général & contexte du projet

Dans ce chapitre, je donnerai un aperçu des différentes technologies utilisées dans ce projet, et je ferai une présentation du contexte du projet.

Généralités

Un DataMart permet d'intégrer des sources de données hétérogènes à des fins d'analyse. Un des points clé de la réussite du processus d'entreposage de données réside dans la définition du modèle de l'entrepôt en fonction des sources de données et des besoins d'analyse. Une fois l'entrepôt conçu, le contenu et la structure des sources de données, tout comme les besoins d'analyse, sont amenés à évoluer et nécessitent ainsi une évolution du modèle de l'entrepôt (schéma et données).

1. Concept d'un DataMart ou entrepôt de données

1.1 Présentation

Le concept d'entrepôt de données a été formalisé pour la première fois en 1990. L'idée de constituer une base de données orientée sujet, intégrée, contenant des informations datées, non volatiles et exclusivement destinées aux processus d'aide à la décision fut dans un premier temps accueillie avec une certaine hésitation. Beaucoup n'y voyaient qu'une autre forme du concept déjà ancien : l'infocentre.

L'entreprise doit anticiper pour faire face aux nouveaux enjeux économiques. Pour être efficace, l'anticipation peut s'appuyer sur de l'information pertinente qui est à la portée de toute entreprise qui dispose d'un capital de données gérées par ses systèmes opérationnels et qui peut en acquérir d'autres auprès de fournisseurs externes. Mais ces données ne sont pas organisées dans une perspective décisionnelle et sont éparpillées dans plusieurs systèmes hétérogènes. Il est nécessaire de rassembler et d'homogénéiser les données afin de permettre des analyses des indicateurs pertinents et de faciliter les prises de décisions.
Pour répondre à ces besoins, il a été défini et intégré une architecture qui va servir de fondation aux applications décisionnelles : l'entrepôt de données.

1.2 Pourquoi un entrepôt de donnée

L'entreprise construit un système décisionnel pour améliorer sa performance. Elle doit décider et anticiper en fonction de l'information disponible et capitaliser sur ses expériences.

Depuis plusieurs dizaines d'années, une importante masse d'informations est stockée sous forme informatique dans les entreprises. Les systèmes d'informations sont destinés à garder la trace d'événements de manière fiable et intègre. Ils automatisent de plus en plus les processus opérationnels.

L'informatique a un rôle à jouer, en permettant à l'entreprise de devenir plus entreprenante et d'avoir une meilleure connaissance de ses clients, de sa compétitivité ou de son environnement.

1.3 La réalité des systèmes d'information

Les données contenues dans les systèmes d'informations sont :

- **Eparpillées :**

  Il existe souvent de multiples systèmes, conçus pour être efficaces pour les fonctions sur lesquelles ils sont spécialisés.

- **Peu structurées pour l'analyse :**

  La plupart des systèmes informatiques actuels ont pour objet de conserver en mémoire l'information, et sont structurés dans ce but.

- **Focalisées pour améliorer le quotidien :**

  Toutes les améliorations technologiques se sont focalisées pour améliorer cette capacité en termes de volume, qualité, rapidité d'accès. Il manque très souvent la capacité à donner les moyens de tirer parti de cette mémoire pour prendre des décisions.

- **Utilisées pour des fonctions critiques :**

  La majorité des systèmes existants est conçue dans le but unique de répondre aux besoins avec des temps de traitement corrects.

Le tableau suivant présente les différences entre les données opérationnelles et décisionnelles.

Données opérationnelles	Données décisionnelles
▪ Orientées application, détaillées, précises au moment de l'accès	▪ Orientée activité (thème, sujet), condensées, représentes des données historiques
▪ Mise à jour interactive possible de la part des utilisateurs	▪ Pas de mise à jour interactive de la part des utilisateurs
▪ Accédées de façon unitaire par une personne à la fois	▪ Utilisées par l'ensemble des analystes, gérées par sous-ensemble
▪ Cohérence atomique	▪ Cohérence globale
▪ Haute disponibilité en continu	▪ Exigence différente, haute disponibilité ponctuelle
▪ Structure statique, contenu variable	▪ Structure flexible
▪ Petite quantité de données utilisées par un traitement	▪ Grande quantité de données utilisée par les traitements
▪ Réalisation des opérations au jour le jour	▪ Cycle de vie différent
▪ Utilisées de façon répétitives	▪ Utilisées de façon aléatoires

1.4 Les objectifs

Toutes les données, qu'elles proviennent du système de production de l'entreprise ou qu'elles soient achetées, vont devoir être organisées, coordonnées, intégrées et stockées, pour donner à l'utilisateur une vue intégrée et orientée métier. L'entrepôt de données doit viser les objectifs suivants :

- Rendre les données de l'organisation facilement accessibles.
  Le contenu de l'entrepôt de données doit être facile à comprendre. Les données doivent être parlantes et leur signification évidente pour l'utilisateur et pas seulement pour le développeur.

- Présenter l'information de l'organisation de manière cohérente.
  Les données de l'entrepôt doivent être crédibles.

- Etre adaptable et résistant aux changements.
  Les données de l'entrepôt doivent être conçues pour traiter les changements. De ce fait, les changements ne doivent pas invalider les données existantes ou les applications.

- Etre le socle sur lequel repose l'amélioration des prises de décision.

- Etre accepté par les utilisateurs pour pouvoir réussir

2. Composants d'un entrepôt de données

2.1 Les applications opérationnelles sources

Ce sont les applications opérationnelles qui capturent les transactions de l'organisation. Les applications sources ne conservent que très peu de données historisées. Un bon entrepôt de données peut libérer les applications sources d'une bonne partie de leurs responsabilités concernant la représentation du passé.

2.2 La préparation de données

La zone de préparation des données de l'entrepôt est à la fois une zone de stockage et un ensemble de processus couramment appelés ETL (Extract/Transform/Load). L'extraction est la première étape du processus d'apport de données à l'entrepôt qui se traduit par la lecture, l'interprétation et la copie des données sources dans la zone de préparation. Ensuite, on passe à la transformation en vue du chargement. Il faut interdire aux utilisateurs l'accès à la zone de préparation des données. La dernière étape s'occupe de charger les données, préalablement extraites puis transformées, dans des cibles hétérogènes (le plus souvent des entrepôts de données).

2.3 La présentation de données

La zone de présentation des données est le lieu où les données sont organisées, stockées et offertes aux requêtes directes des utilisateurs, aux programmes de reporting et aux autres applications d'analyse. La zone de présentation des données est l'entrepôt de données tel qu'il est vu par les utilisateurs.

2.4 Les outils d'accès aux données

L'ensemble des outils d'accès aux données constitue le dernier composant majeur d'un environnement d'entrepôt de données. Les outils d'accès aux données constituent l'ensemble des moyens fournis aux utilisateurs pour exploiter la zone de présentation en vue de prendre des décisions basées sur des analyses.

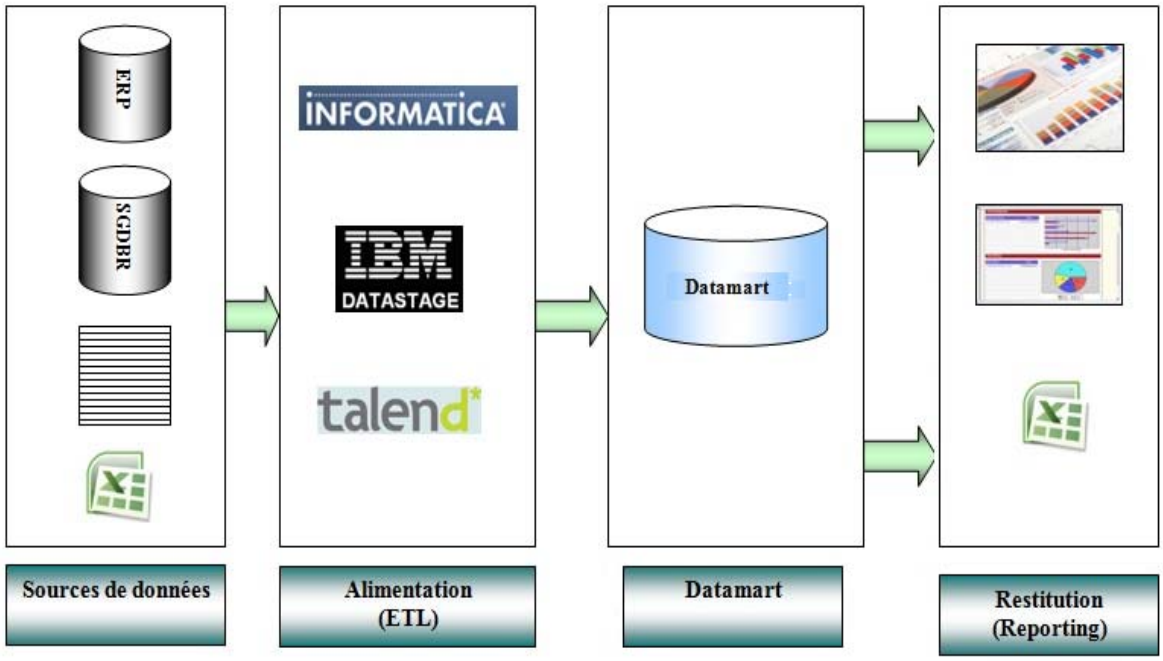


Figure 5 : ARCHITECTURE GENERALE D'UN ENTREPOT DE DONNEES

3. Le décisionnel et les logiciels libres

Les projets décisionnels sont divisés, le plus souvent, en deux phases. La première consiste en extraction, la transformation et l'alimentation des données et elle nécessite l'utilisation d'outils de type ETL (DataStage ou Talend). Les outils d'alimentation présents sur le marché sont multiples et variés. Ils se présentent sous formes libres ou propriétaires et nécessitent donc une étude préalable afin de déterminer, en fonction du travail à réaliser, l'outil adéquat.

La seconde phase est celle du reporting. Elle va produire les rapports de synthèse finaux qui vont permettre d'offrir une vue globale de l'activité. Il existe peu de logiciels de reporting libres sur le marché (Jasper Soft). Au sein d'Orange Labs, on trouve l'outil propriétaire Business Object.

Pour mieux comprendre le déroulement d'un projet décisionnel intégrant des outils libres (open source), je vais commencer par une brève introduction sur les logiciels libres et expliquer le fonctionnement des ETL avant de terminer par une présentation des 2 outils décisionnels utilisés par l'URD ISA d'Orange Labs à Belfort.

3.1 Les logiciels libres

Il existe de nombreux outils décisionnels propriétaires sur le marché actuel dont le coût est très élevé. Dans tous les domaines, on observe l'intérêt des entreprises pour les solutions libres qui sont moins onéreuses ; mais plusieurs interrogations se posent sur leurs performances et leur robustesse (problème de la maintenance et l'absence, en général, de support)

Les fonctionnalités des logiciels propriétaires sont déterminées par la demande générale des entreprises, ce qui ne satisfait pas véritablement tous les utilisateurs. En revanche, les solutions libres vont permettre à une entreprise de moduler une application en se basant sur ses propres besoins et non sur ceux du marché. Cette démarche est réalisée par de grandes entreprises comme par des PME qui s'adressent en général à des SSII (Société de Services en Ingénierie Informatique) afin de développer des fonctionnalités qui s'adaptent à leurs besoins.

Une question importante se pose sur la redistribution de ces logiciels par ces entreprises car au final peu sont celles qui suivent la logique du libre, qui est avant tout de partager toute amélioration d'un logiciel libre.

Certaines entreprises sont réticentes face aux logiciels libres malgré leur notoriété et leurs coûts abordables. Les principaux facteurs de blocage qui ralentissent encore la diffusion des solutions libres sont entre autres :

- L'immaturité des solutions
- La sécurité et les inquiétudes liées à l'existence de "communautés"
- Les défaillances de l'assistance
- L'absence de fournisseurs dominants
- Les craintes liées aux licences logicielles
- Les performances et la richesse fonctionnelle des solutions

3.2 Le fonctionnement des ETL

Un ETL (**Extract – Transform – Load**) permet l'Extraction, la Transformation et le Chargement de données depuis des sources diverses (bases de données, fichiers) vers des cibles préalablement définies comme le montre le schéma ci-dessous.

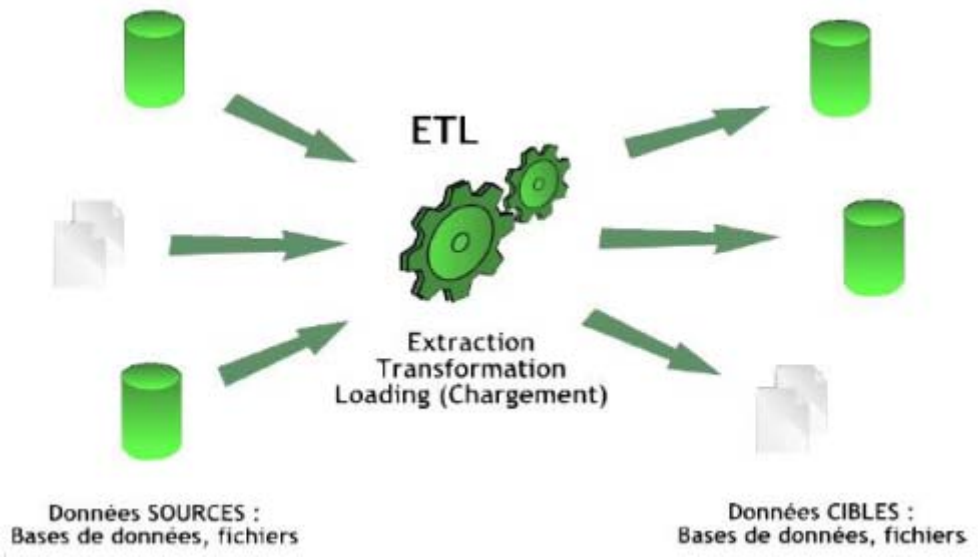


Figure 6 : SCHEMA DE FONCTIONNEMENT D'UN ETL

Les ETL sont communément utilisés dans l'informatique décisionnelle afin de permettre l'alimentation des bases de données décisionnelles (Datawarehouse, DataMart). Ces dernières servent de supports pour l'analyse des données sous plusieurs formes :

- Rapports et états
- Tableaux de bords
- Indicateurs de performance
- Analyse multidimensionnelle
- Analyse exploratoire

Les volumes de données traités sont plus ou moins importants. Ainsi, les critères de performance sont primordiaux dans le choix d'un ETL.

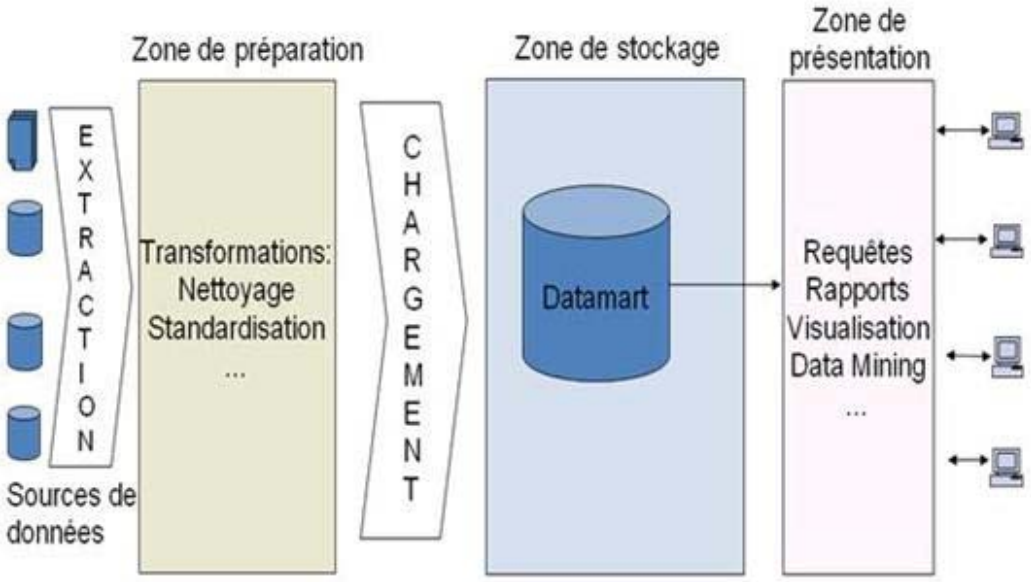


Figure 7 : SCHEMA REPRESENTANT LA CHAÎNE DECISIONNELLE

Le choix le plus difficile dans tout projet décisionnel ou d'intégration/migration de données consiste à déterminer quelle méthode doit être mise en œuvre :

1. Faut-il créer du code spécifique (procédures SQL, code Java ou autre) ?

2. Faut-il acheter un ETL propriétaire (DataStage, Informatica ou autre) ?

La première solution semble intéressante, car elle permet de rester au plus près des spécificités métiers des données à traiter, tout en s'affranchissant des contraintes liées à l'achat et l'utilisation d'un ETL propriétaire. Cependant, cette solution peut s'avérer coûteuse à long terme, tout simplement car l'évolutivité constante des données métier entraîne une nécessaire adaptation des traitements d'intégration. Celle-ci n'est pas toujours facile à gérer, surtout si les équipes projets évoluent au cours du temps.

La deuxième solution va permettre de mettre en œuvre très rapidement les traitements d'intégration, avec cependant des coûts élevés (achat des licences, formations, maintenance...) et ceci dès la phase de démarrage du projet.

Il existe désormais une solution alternative : **utiliser un ETL libre.**

On bénéficie ainsi des avantages d'un ETL tout en gardant une maîtrise lissée des coûts.

Ces derniers sont en effet réduits aux coûts de formation initiale de l'outil et d'une éventuelle souscription à une hotline technique. Aucune licence n'est à payer dans ce modèle économique.

Les ETL libres qui paraissent à l'heure actuelle comme les plus intéressants en termes de fonctionnalités proposées, de maturité et de pérennité sont entre autre Pentaho Data

Integration et Talend Open Studio. Ils sont en mesure de répondre de façon équivalente à la plupart des ETL propriétaires disponibles sur le marché.

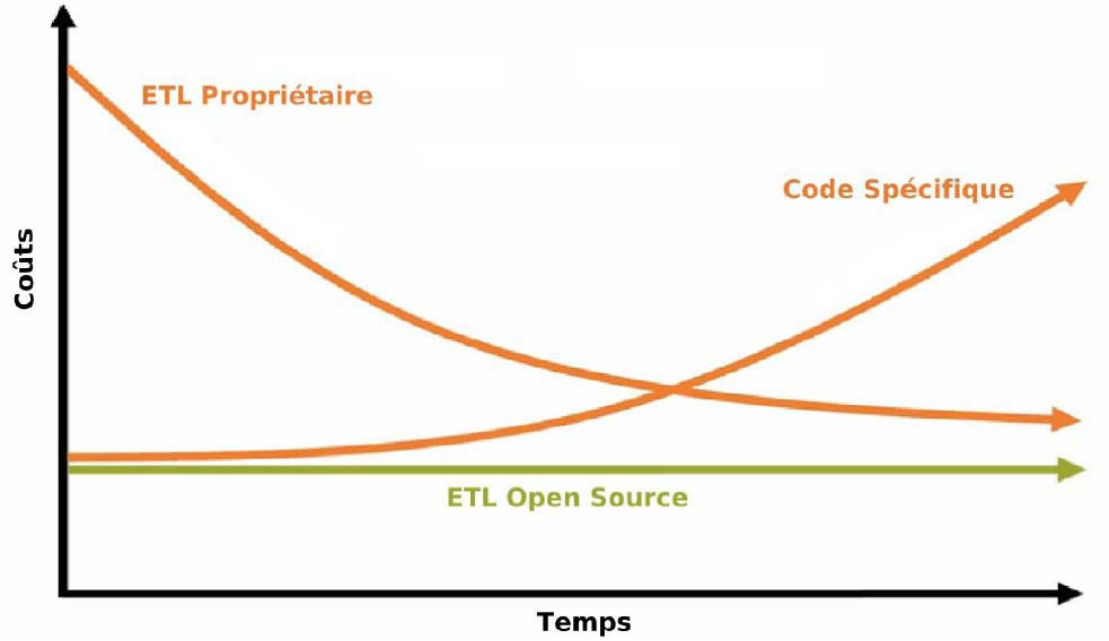


Figure 8 : COUTS EN FONCTION DU TEMPS POUR LES DIFFERENTES SOLUTIONS

3.3 Les outils ETL

Cette partie permet de mieux comprendre le fonctionnement de l'ETL libre Talend Open Studio et l'ETL propriétaire DataStage de la société IBM.

3.3.1 Talend Open Studio (TOS)

Talend Open Studio est développé par Talend, une société française dynamique et relativement jeune. La première version de « Talend Open Studio » a vu le jour au 2ème semestre 2006, et la version actuelle est la 5.1.0

Talend Open Studio est un ETL du type « générateur de code ». Pour chaque traitement d'intégration de données, un code spécifique est généré en Java. Les données traitées et les traitements effectués sont donc intimement liés.

Talend Open Studio utilise une interface graphique, le « Job Designer » (basée sur Eclipse RCP) qui permet la création des processus de manipulation de données.

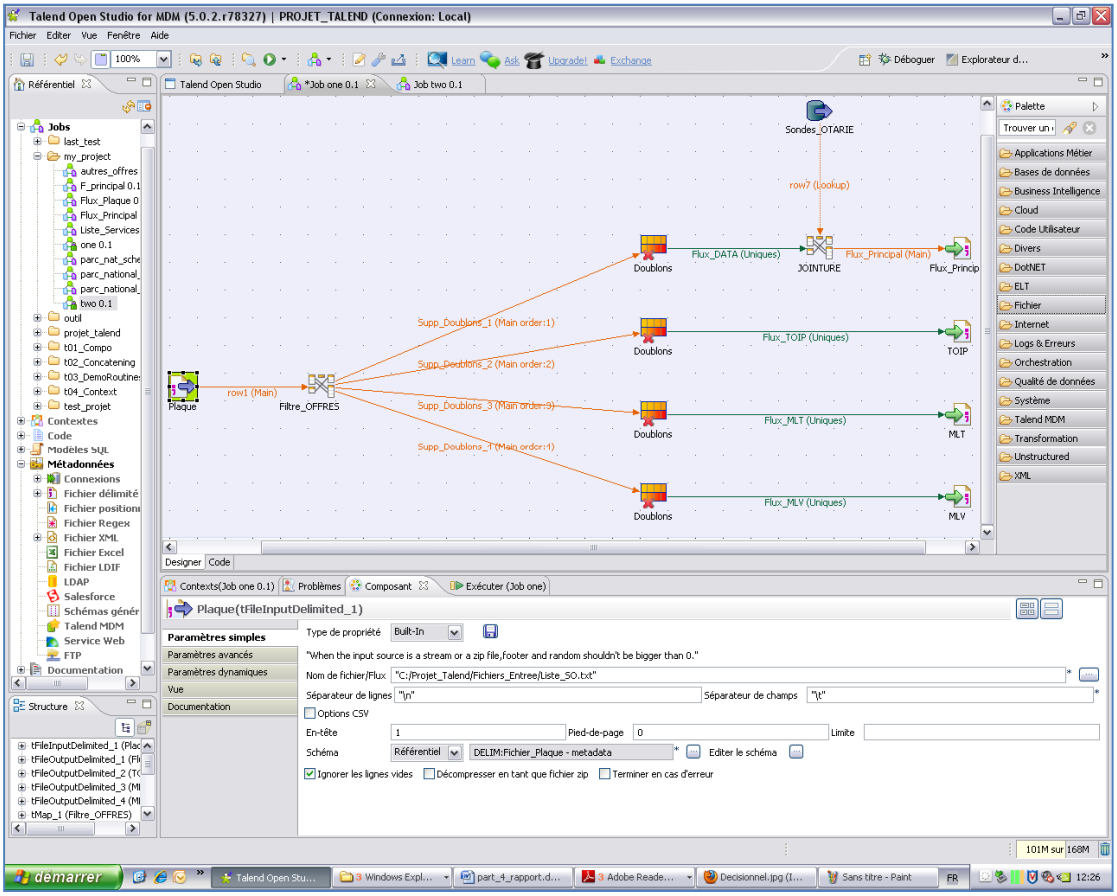


Figure 9 : ZONE DE TRAVAIL DU LOGICIEL TALEND

De nombreux types de composants sont disponibles pour se connecter aux principaux SGBD (Oracle, DB2, MS SQL Server, PostgreSQL, MySQL,...) ainsi que pour traiter tous les types de fichiers plats (CSV, Text, XML), aussi bien en lecture qu'en écriture. Talend facilite la construction des requêtes dans les bases de données en détectant le schéma et les relations entre tables.

Un référentiel permet de stocker les métadonnées afin de pouvoir les exploiter dans différents jobs. Par exemple, on peut sauvegarder le type et le format des données d'entrée d'un fichier CSV afin de pouvoir les exploiter ultérieurement dans un ou plusieurs composants, facilitant ainsi toute évolution éventuelle du schéma.

La conception très visuelle des traitements permet de présenter des statistiques d'exécution en temps réel ou encore de tracer les données transitant ligne à ligne dans les composants de la chaîne de traitement. Quand un job d'intégration est lancé via le Job Designer (en mode graphique), il est possible d'afficher les statistiques de traitement en temps réel, montrant le nombre de lignes traitées et rejetées, ainsi que la vitesse d'exécution (lignes par secondes).

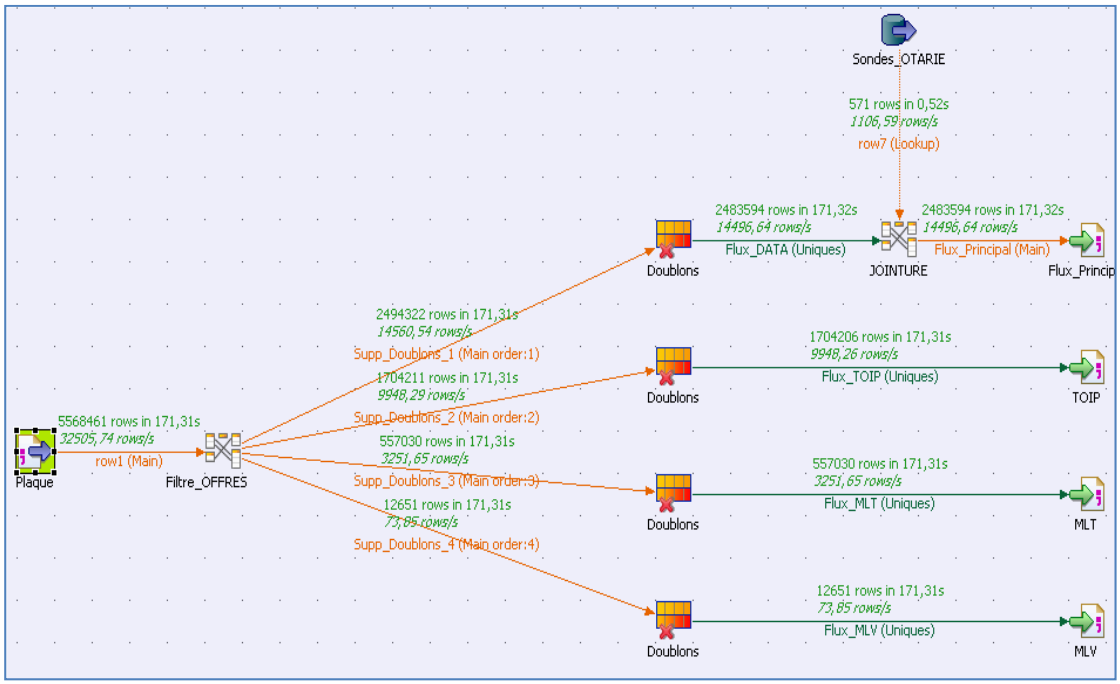


Figure 10 : EXEMPLE DE TRANSFORMATION SOUS TALEND 5.0

3.3.2 IBM InfoSphere DataStage

IBM InfoSphere DataStage permet de collecter, d'intégrer et de transformer de gros volumes de données, quelle que soit la complexité des structures. Il permet d'intégrer toutes les informations de l'entreprise, quels que soient le nombre de sources/cibles et les délais. Qu'il s'agisse de créer un entrepôt de données répondant aux besoins informationnels de l'entreprise – et ce éventuellement en temps réel - ou d'intégrer plusieurs dizaines de systèmes source prenant en charge les applications d'entreprise, telles que les applications de gestion de la relation client et de gestion de la chaîne logistique globale ou les applications ERP (Enterprise Resource Planning), IBM InfoSphere DataStage garantit des informations fiables.

IBM InfoSphere DataStage offre trois avantages décisifs pour le succès de l'intégration des données des entreprise : une connectivité très complète pour accéder facilement et rapidement à n'importe quel système source ou cible ; des outils de développement et de maintenance avancés qui accélèrent la mise en œuvre et simplifient la gestion des données ; et enfin une plateforme évolutive qui facilite le traitement des données de l'entreprise.

IBM InfoSphere DataStage accepte un nombre pratiquement illimité de sources et de cibles de données pour une même tâche :

- Fichiers texte
- Structures de données XML complexes
- Systèmes d'applications d'entreprise tels que SAP, Siebel, Oracle et PeopleSoft
- Quasiment toutes les bases de données, y compris les bases de données partitionnées, telles qu'Oracle, IBM DB2 Universal Database, IBM Informix, Sybase, Teradata et Microsoft SQL Server
- Services Web
- SAS
- Produits d'intégration d'applications d'entreprise et de messagerie, tels que WebSphere MQ et SeeBeyond, etc.

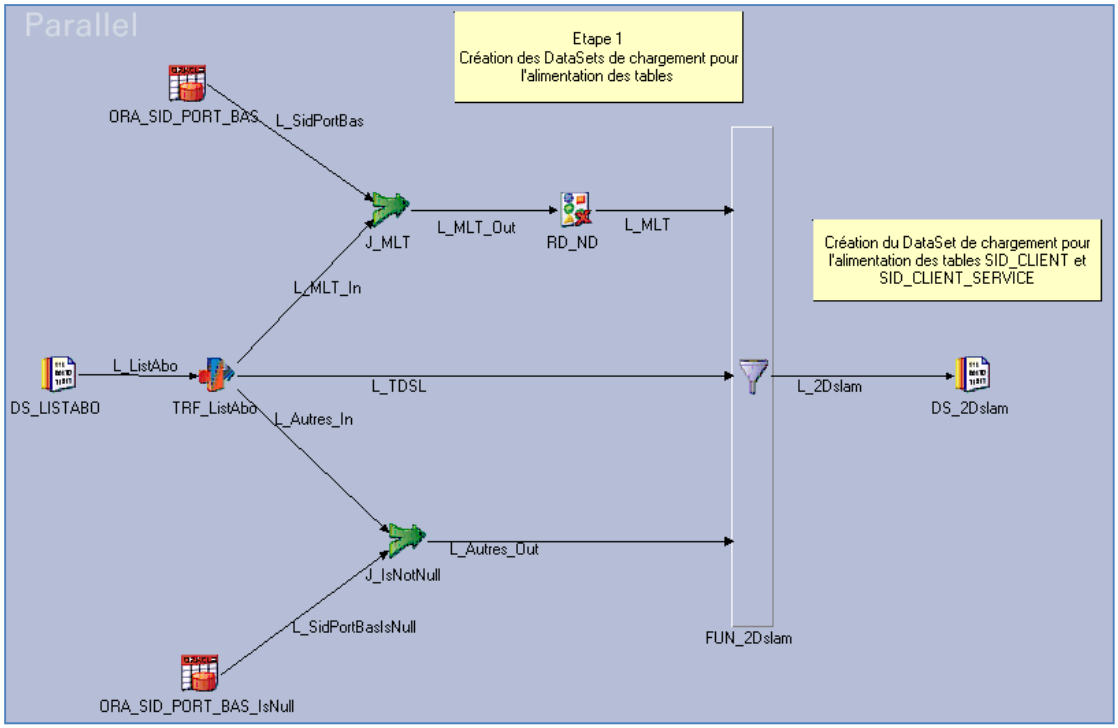


Figure 11 : EXEMPLE DE TRANSFORMATION SOUS DATASTAGE

Contexte du stage

1. Présentation

Le DataMart est un ensemble de données ciblées et organisées pour répondre à un besoin spécifique à un domaine ; celui traité dans ce document concerne les clients ADSL d'Orange. Ce magasin de données est destiné à être interrogé et à fournir un panel d'informations permettant d'améliorer les services et le trafic des clients. Pour ce faire, une étude de l'existant est indispensable car elle va permettre de définir les différents axes d'amélioration sur lesquels vont porter les rapports à fournir.

2. Objectifs

De façon plus technique, ce DataMart a pour vocation de présenter les données de manière spécialisée, agrégée et regroupée fonctionnellement. Il permet de restituer tout le spectre de l'activité des clients ADSL sous forme d'étapes de travail ou de rapports de synthèse.

L'objectif de cette étude est de fournir dans un premier temps un prototype sur la base de solutions décisionnelles pour tester l'ETL open source Talend sur un cas réel afin d'en dégager le potentiel et d'en définir éventuellement les coûts de migration.

Ensuite, à partir des données d'Orange obtenues de différentes sources (SI, sondes), on effectue un ensemble d'opérations. Cela consiste à combiner, agréger et croiser ces données pour obtenir une collection de données cohérente et structurée, basée sur des contraintes techniques, fonctionnelles et économiques, en vue de répondre aux besoins de la MOA.

3. Description

Analyser et prévoir les différents types de trafics des clients d'Orange est essentiel afin d'élaborer des offres adaptées au mieux à leur demande et afin d'anticiper les évolutions des réseaux nécessaires pour supporter ces nouvelles offres et garantir leurs qualités. C'est pourquoi RESA/PEAK/ISA développe un DataMart permettant un meilleur suivi des trajectoires des trafics ADSL par la production d'indicateurs clés facilitant la réalisation des études de trafics et alimentant différentes entités réseau et marketing du groupe.

Durant les 5 mois passés au sein d'Orange Labs avec l'équipe ISA, j'avais pour première mission d'étudier et d'analyser la chaîne de traitement qui fonctionne avec l'ETL DataStage. L'objectif était la montée en compétence sur le domaine d'étude et

sur l'ETL et la rédaction d'un document technique décrivant les étapes de travail ainsi que la structure du schéma fonctionnel.

En me basant sur les résultats de la première partie, la deuxième mission était l'analyse et la conception d'une chaîne de traitement plus rationnelle. L'objectif final était d'effectuer le développement sous l'ETL open source Talend Open Studio.

Chapitre 3 Description synthétique du projet

Dans ce chapitre, je donnerai un aperçu sur les différents volets du projet, en détaillant plus amplement les exigences et les besoins du groupe Orange au niveau de chaque partie du projet.

Projet Network Mining du groupe Orange

1. Présentation

Les clients d'Orange se voient aujourd'hui proposer des offres de très haut débit, que ce soit sur leur mobile ou sur internet, leur permettant ainsi d'apprécier des services multiples comme la visiophonie, la télévision et la vidéo à la demande avec une haute définition. On tend également vers une convergence des moyens d'accéder au réseau Orange qui est aujourd'hui un opérateur intégré, proposant des offres spécifiques. Dans ce contexte, observer, analyser et prévoir les différents trafics des clients fixes et mobiles, voix et data, est essentiel, pour proposer des offres de qualité et adaptées aux besoins des clients, tout en mesurant l'impact de ces nouvelles offres au niveau du réseau. C'est pourquoi FT/OLNC/RD/RESA/PEAK/ISA développe un DataMart pour aider à l'analyse des trafics de ses clients. Ce DataMart est alimenté par des données provenant du système d'information d'Orange et par des données de trafic provenant de l'application OTARIE développée par l'unité Recherche & Développement qui récolte les informations des sondes disposées sur le réseau.

2. Cahier des charges

Pour la construction du DataMart, il s'agit de développer toutes les briques logicielles et le système de base de données pour contenir le gisement :

- Construire la structure de données Oracle pour accueillir les données (modélisation de la base d'intégration, création de scripts et création des tables)

- Télécharger les données depuis le SI ou depuis le serveur de collecte des captures de trafic.

- Développer les traitements ETL

  - ✓ Pour transformer, croiser, agréger les données sources et alimenter ainsi la base d'intégration.
  - ✓ Pour extraire les données pertinentes et alimenter la base de diffusion

- Réaliser un reporting au format de fichiers texte

- Tester et valider les données et les indicateurs extraits.

2.1 Objectifs

Les objectifs sont multiples ; ils consistent à :

- Donner un accès rapide et simple à l'information stratégique à la MOA (Maître d'ouvrage)
- Rafraichir et mettre à jour les documents fournis à partir des activités des clients ADSL
- Développer des services qui satisfont les besoins des clients d'Orange
- Répondre aux questions marketing
- Mettre en place un système d'information dédié aux applications décisionnelles.

3. Présentation du projet :

Dans cette partie, je présente le DataMart mis en œuvre à FT/OLNC/RD/RESA/PEAK/ISA afin de collecter les informations issues du trafic des clients ADSL d'Orange pour pouvoir les analyser et fournir des informations à la MOA. Tous les traitements sont effectués par l'ETL DataStage (IBM), que ce soit pour alimenter les différentes tables de la base de données, ou pour fournir des informations sous forme de fichiers texte. Tous ces traitements sont effectués par des jobs successifs, pouvant être exécutés par un séquenceur.

Le lancement de ces Jobs peut être effectué au moyen de scripts bash. Les formats d'entrée et de sortie de cet outil sont très variés :

- Base de données
- Fichiers texte
- Fichiers Hash Files (fichiers Pick Universe)

3.1 Description fonctionnelle

A partir d'un gisement de données (donnée brutes) provenant de différentes sources, des données descriptives du réseau mais aussi des données de capture du trafic, il s'agit de modéliser une base d'intégration en permettant leur croisement. Les données ont été traitées et normalisées d'une manière cohérente et organisée sous forme d'un SGBD orienté Décisionnel.

Une base de diffusion est ensuite obtenue à partir de la base d'intégration. Cette base de diffusion contiendra des données agrégées, croisées et historisées permettant aux statisticiens d'avoir immédiatement les informations dont ils ont besoin. Cette base est organisée de telle sorte qu'elle soit accessible par plusieurs utilisateurs.

L'objectif est de permettre une réponse rapide à toutes les problématiques posées par le dimensionnement et la qualité de service des réseaux de collecte ADSL.

3.2 Sources des données

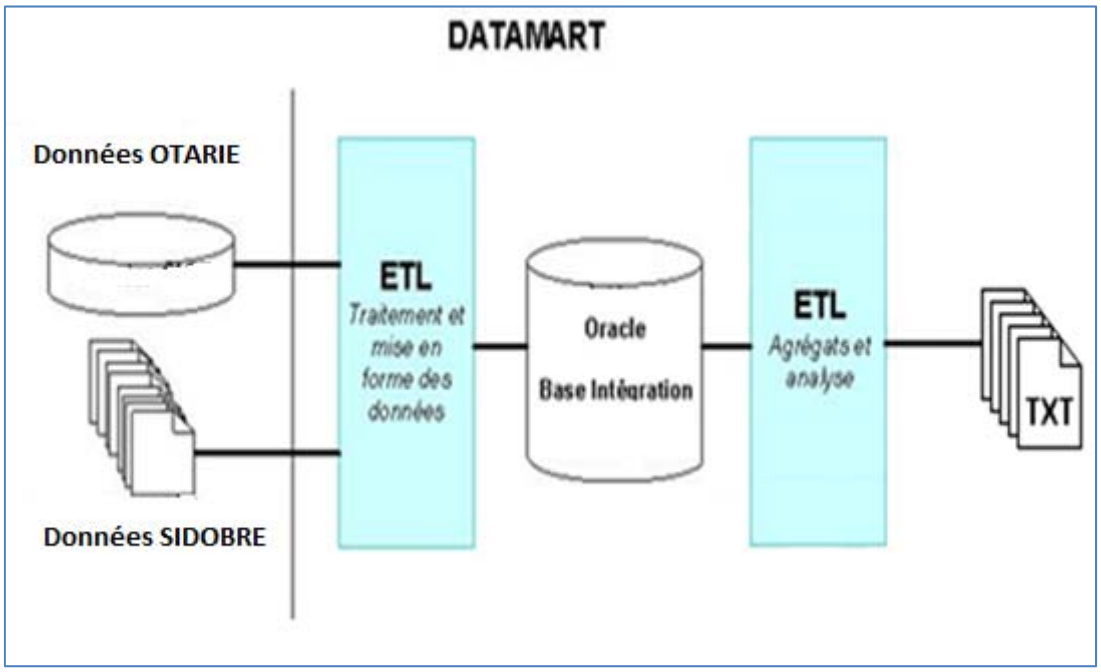


Figure 12 : ARCHITECTURE FONCTIONNELLE DU DATAMART TRAFIC ADSL

Le DataMart trafic utilise d'une part les données internes, journalières, issues de l'application SIDOBRE, contenant les caractéristiques techniques des clients qui ont généré du trafic ADSL et leur offre associée, et d'autre part celles externes issues de l'application OTARIE qui collecte les données de trafic.

Les données SIDOBRE sont mises à disposition 7j/7 via un portail web. Quant aux données OTARIE, elles sont accessibles sur un serveur situé à Lannion.

3.3 Architecture technique

Le DataMart est réparti sur 4 serveurs reliés entre eux via le réseau Ermes, sur lequel sont raccordées toutes les machines d'Orange Labs. Les serveurs sont les suivants :

NOM	LIEU	Système d'exploitation	DISQUE
P-NSOISE	Issy-les-Moulineaux	Windows Server 2003	Data : 585 Go Data2 : 683 Go
B-DIBUS	Belfort	Windows Server 2003	Data : 817 Go Data2 : 817 Go
B-TRAFIC	Belfort	Linux RedHat RHEL 4 x86 32 bits	/data/data1 :2,0 To /data/data2 : 1,8 To
B-NTWMINING	Belfort	Linux RedHat RHEL 4 x86 32 bits	/srv/nas : 2,2 To

L'architecture technique est la suivante :

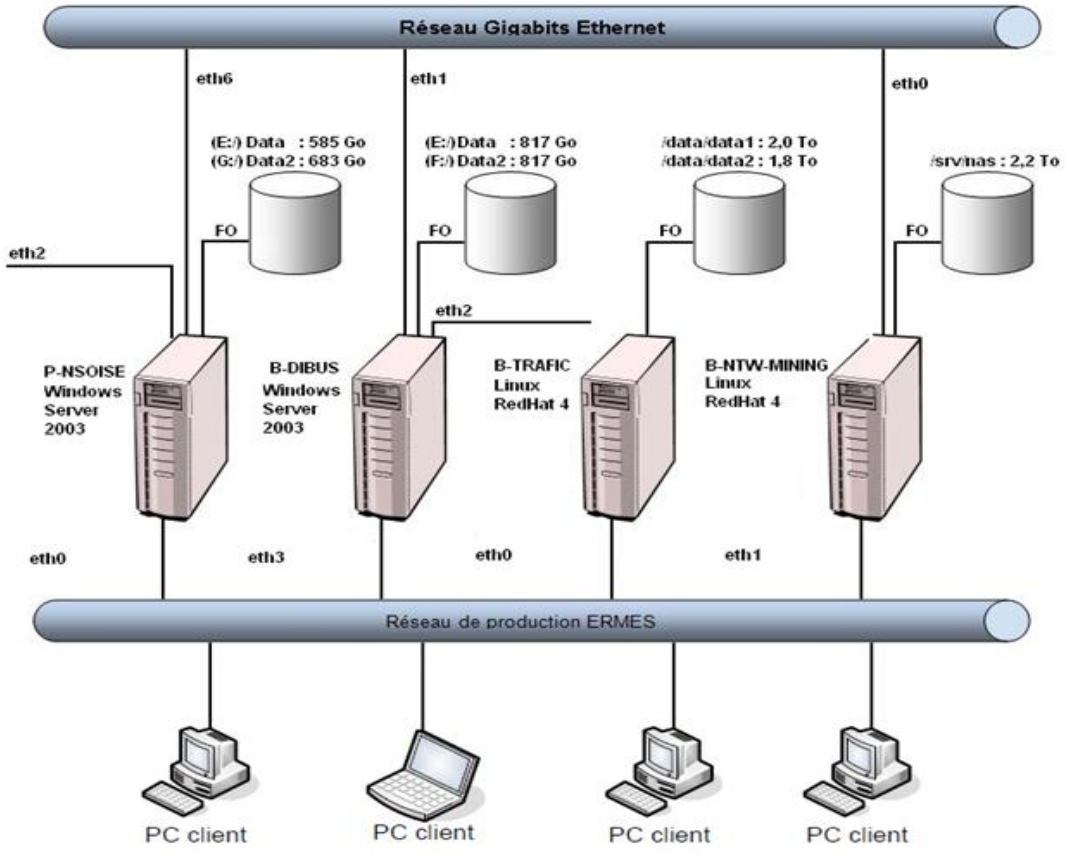


Figure 13 : L'ARCHITECTURE TECHNIQUE DES SERVEURS DATAMART

4. La chaîne de traitement du DataMart Trafic ADSL

La chaîne de traitement actuellement utilisée est sur le serveur B-TRAFIC avec la version DataStage 7.5.3 pour Linux (Redhat 4). Cette chaîne permet d'alimenter la base de données Oracle installée sur le serveur B-DIBUS. Cette base contient le schéma SIDOBRE qui est constitué de 13 tables dont 2 seront mises à jour par la chaîne de traitement étudiée :

- **SID_CLIENT :** informations générales sur les clients
- **SID_CLIENT_OTARIE_CATM** : liste des clients ATM par journée de trafic
- **SID_CLIENT_OTARIE_CGE** : liste des clients GE par journée de trafic
- **SID_CLIENT_SERVICE** : identification et informations techniques des clients
- **SID_COUPLE_OFF_SCE** : identifications de l'offre client
- **SID_DSLAM** : liste des DSLAM
- **SID_HISTO_BAS** : dates de changement de BAS des clients (mutation)
- **SID_HISTO_CLIENT** : permet l'historisation des clients
- **SID_OFFRE** : correspondance Offre SIDOBRE/Offre OTARIE
- **SID_OFFRE_COM** : correspondance Offre SIDOBRE/Liste des services
- **SID_PARC** : indication du nombre de clients (jour, BAS et offre OTARIE)
- **SID_PARC_ECR** : indication du nombre de clients par jour et par BAS ECR
- **SID_PORT_BAS** : liste des BAS

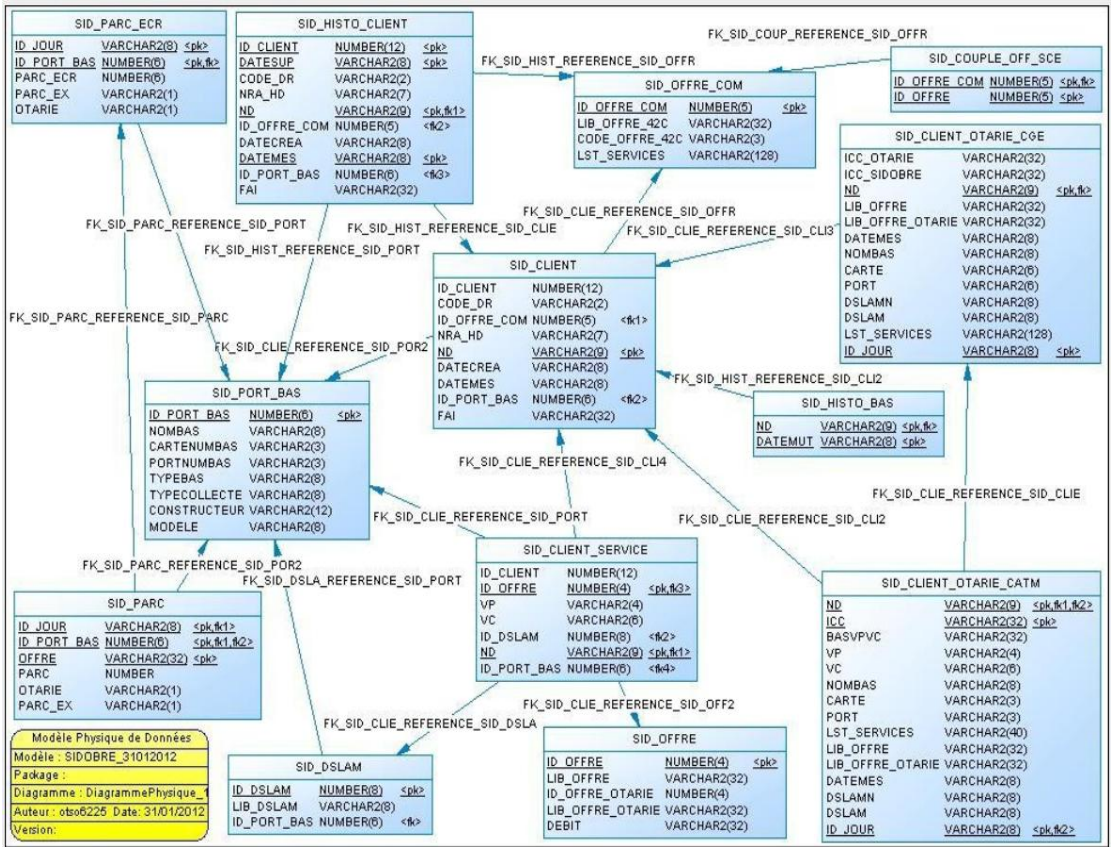


Figure 14 : MODELE PHYSIQUE DES DONNEES SIDOBRE DE LA BASE ORACLE TRAFIC

Les données SIDOBRE proviennent du système d'information d'Orange et regroupent des informations sur les clients d'Orange, leurs offres associées et leurs mises en œuvre technique (Nom du BAS, VP et VC du client…). Les données sont contenues dans des fichiers journaliers, permettant une mise à jour rapide des clients dans le DataMart. Un client est identifié par son numéro de désignation (numéro de téléphone).

Les fichiers SIDOBRE sont mis à disposition sur un portail Web. Ceux récupérés dans le cadre du DataMart le sont tous les jours (sauf le dimanche), en fin de matinée :

- 5 fichiers de parc des clients ADSL des 5 plaques du territoire français
- 2 fichiers d'occupation des BAS OTARIE
- 13 fichiers FTTH **(**non utilisés dans nos traitements)

Les informations issues des 5 fichiers de plaque contenant la liste des clients ayant généré du trafic sont insérées dans le schéma SIDOBRE de la base TRAFIC.

La chaîne de traitement doit fonctionner 6 jours sur 7. Elle est donc exécutée automatiquement sur le serveur B-TRAFIC, à 13H30, à l'aide d'un script lancé depuis la crontab.

Le traitement DataStage se compose d'un séquenceur qui appelle successivement des jobs. Les jobs sont soit des jobs DataStage Server (utilisant le moteur Universe), soit des jobs Parallel Extender (PX) autorisant le parallélisme.

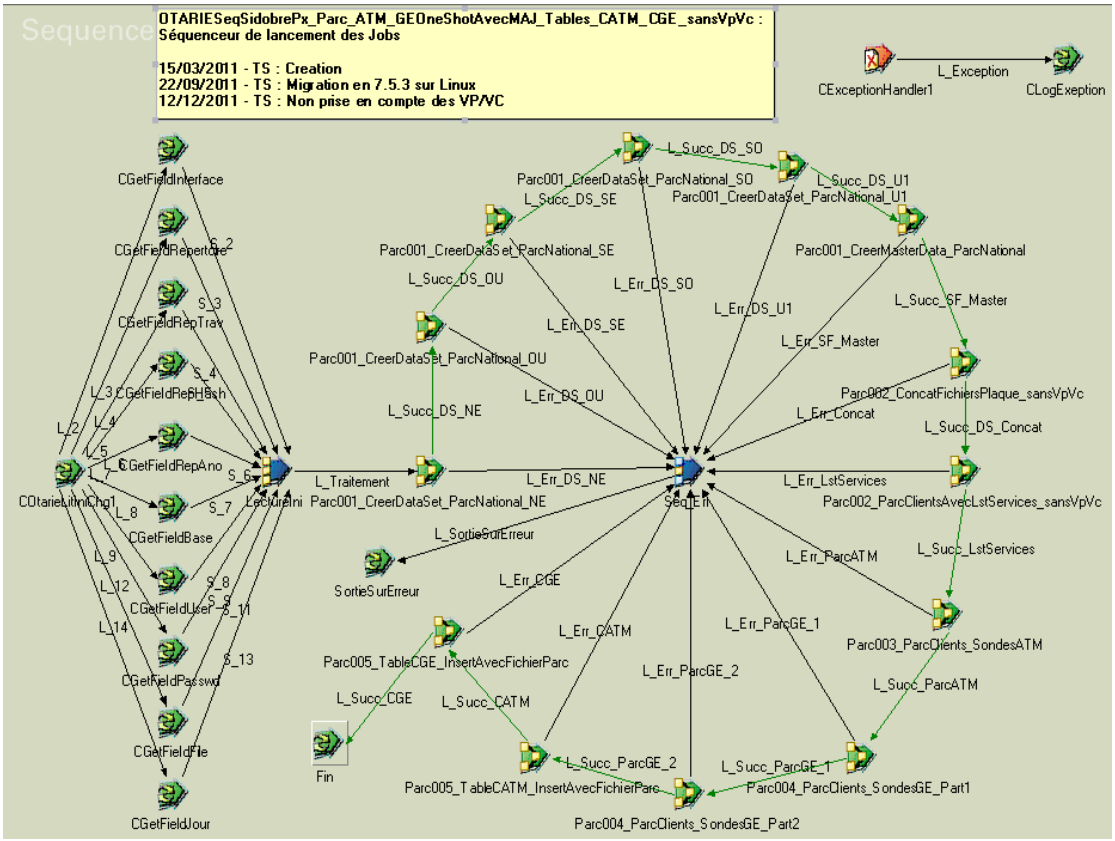


Figure 15 : APPEL DES JOBS VIA UN SEQUENCEUR

Chapitre 4 Présentation du travail réalisé

Dans ce chapitre, je présente mon travail individuel et ma contribution en tant que stagiaire dans ce projet, à savoir, le reverse engineering de l'infocentre existant, l'analyse, la conception et le développement de la chaîne de traitement sous Talend.

Analyse de l'existant

Durant mon stage, j'ai réalisé un reverse engineering du DataMart existant développé sous l'ETL Data Stage de la société IBM. Cela a consisté à décrire la chaîne de traitement dans le but d'avoir une vue globale de son fonctionnement, de comprendre la philosophie d'un ETL, et de voir les données manipulées. Par conséquent, cela m'a permis de monter en compétences sur le domaine de couverture de mon stage.

1. Reverse engineering

Depuis la première version du DataMart trafic ADSL développée il y a plusieurs années, de nombreuses modifications ou évolutions ont été effectuées, avec une mise à jour partielle de la documentation. Par conséquent, j'ai réalisé un reverse engineering afin de décrire la chaîne de traitement dans un document.

L'intérêt de cette opération était triple. Cela m'a permis de prendre en main l'ETL DataStage, de monter en compétence sur les données manipulées et de faire une description de la chaîne existante. Il s'agira dans un premier temps de sélectionner les informations en adéquation avec les objectifs fixés et dans un second temps de déterminer les données à regrouper. Les données issues des sondes existent sous plusieurs formes. Il va donc falloir les intégrer afin de les homogénéiser et de leur donner un sens unique et compréhensible. Il faut avoir une vue globale sur le déroulement du traitement qui s'articule comme suit :

- L'hétérogénéité des sources : des données descriptives du réseau mais aussi des données de capture du trafic
- L'outil existant utilisé par Orange Labs est un outil propriétaire de la société IBM qui se compose d'une architecture client/server, le serveur étant sous linux Redhat.
- Les données sont traitées et normalisées d'une manière cohérente et organisée sous forme d'un SGBD Oracle orienté Décisionnel.

- La chaîne de traitement alimente la base de données et génère des fichiers au format TXT (texte).
- Les données sont agrégées, croisées et historisées pour permettre aux statisticiens d'avoir immédiatement les informations dont ils ont besoin
- Les traitements sont effectués par des jobs successifs, pouvant être exécutés par un séquenceur. Le lancement de ces jobs peut être effectué au moyen d'un script.

1.1 Schéma descriptif

Le schéma global de la chaîne de traitement permet de faire ressortir les éléments permettant de cibler les améliorations possibles sur le DataMart, en vue d'obtenir des gains au niveau du temps de traitement et de la structure des données.

Ce schéma présente le traitement global des données. Les entrées sont les 5 fichiers de plaques, et une table de référence de la base de données Oracle.

Suite à un ensemble d'opérations (concaténation, suppression des doublons, jointure, transformation …), on obtient en sortie un fichier texte contenant le parc national des clients SIDOBRE (ayant eu du trafic ADSL sur la journée de traitement) sur port OTARIE, avec l'ensemble des informations relatives aux clients ainsi que leur offre associée (DATA, TOIP, MLT, MLV).

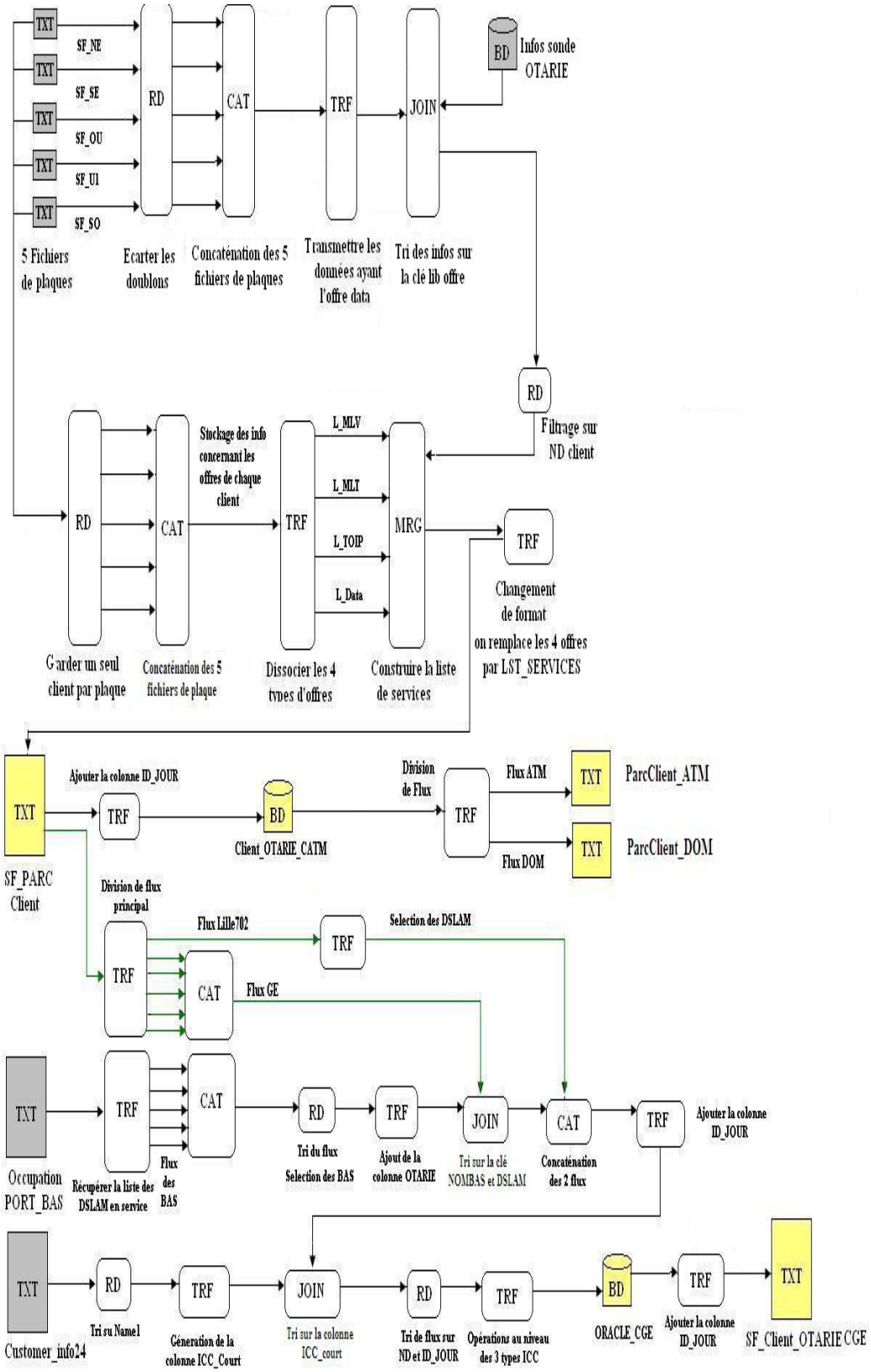


Figure 16 : CHAÎNE DE TRAITEMENT GLOBALE

2. Constat suite au reverse engineering

2.1 Problématique

Les problèmes pour mener à bien ce projet sont nombreux et de natures diverses :

- **Traitement lourd :** un gisement de données (donnée brutes) provenant de différentes sources de nature technique et commerciale.
- **Opérations redondantes ou inutiles** : la chaîne de traitement contient des données redondantes ou non utilisées qui surchargent le flux.
    - Certaines tables d'interrogation sont utilisées dans différents endroits, avec le même traitement qui se répète.
    - la relecture des données sur différentes étapes de travail (fichiers de plaques).
    - Un transfert inutile des données vers d'autres opérations
    - Une clé de filtrage complexe et inadaptée.
    - Non respect des tailles des champs préconisées.
    - Utilisation alternative de plusieurs identifiants pour chercher un élément (exemple : EPC, ND, ID Client)
- **Coût élevé :** pour la réalisation de ces traitements, on se base sur un outil ETL propriétaire avec un coût de maintenance annuel élevé.

2.2 Les axes d'amélioration

Lors de ces traitements, on constate qu'il y a des opérations redondantes ou des données inutiles qui surchargent le flux de données.

Afin de cibler les améliorations possibles et de rationaliser le traitement, il s'avère important de perfectionner la capacité du DataMart en termes de volume, quantité, rapidité d'accès pour prendre les bonnes décisions au bon moment.

L'amélioration de la chaîne de traitement porte sur le changement de la structure du schéma global par l'ajout, la modification et la suppression de certaines opérations :

- Mutualiser les traitements étant auparavant effectués à différents endroits
- Réduire la taille de la clé de filtrage
- Utiliser des mémoires de stockage (éviter l'envoi de plusieurs requêtes à la base de données)
- Enlever les colonnes inutiles
- Utiliser un ETL open source pour test

Mise en œuvre d'un ETL open source

La structuration et le stockage des données dans un entrepôt constituent un support efficace pour permettre des analyses et un développement correct en vue de prises de bonnes décisions.

1. Conception de la chaîne de traitement

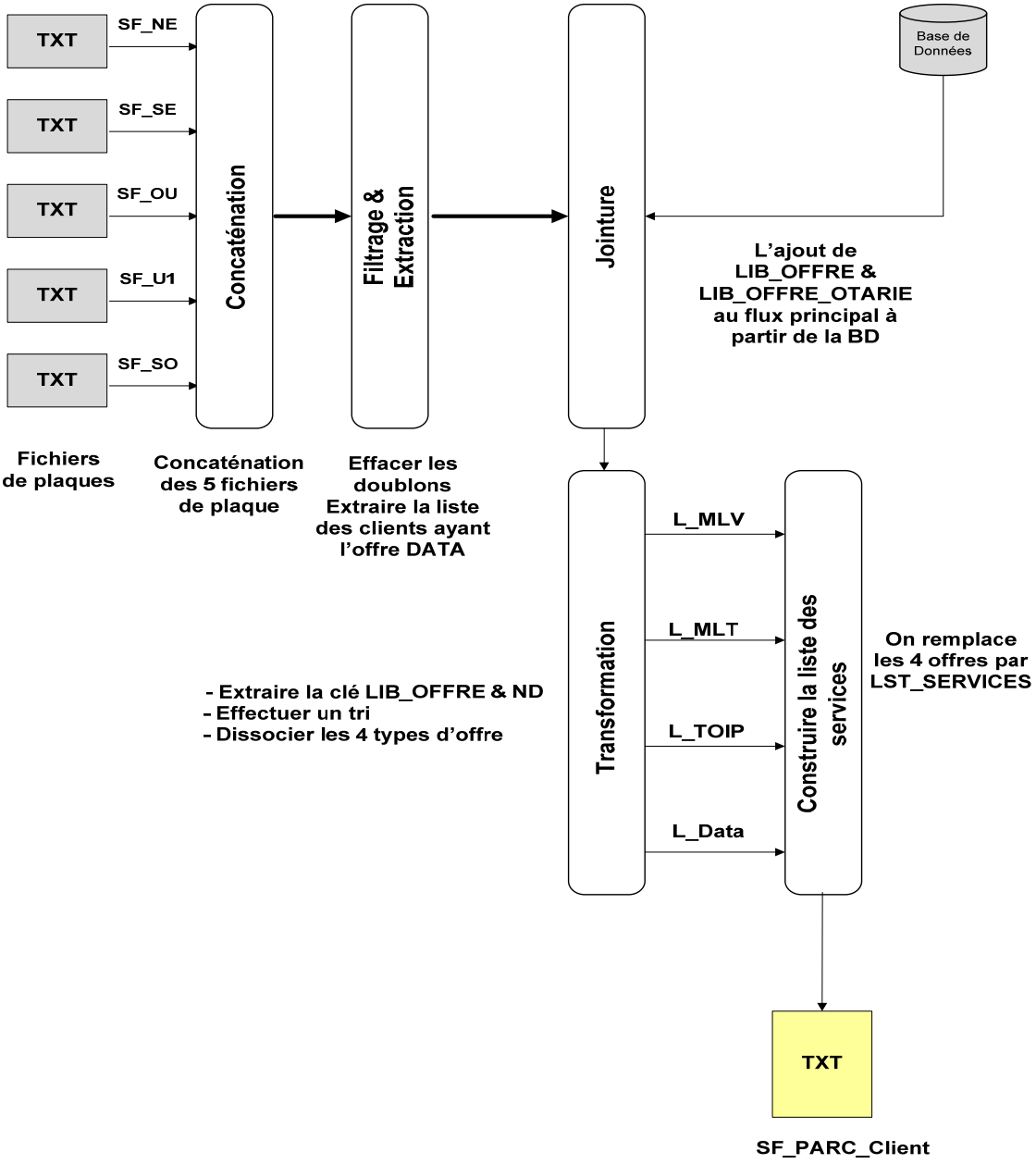


Figure 17 : SCHEMA CIBLE DE LA NOUVELLE CHAÎNE DE TRAITEMENT

Ce schéma présente une amélioration par rapport au schéma de traitement développé sous la chaîne existante au niveau de :

- La réduction du nombre de job
- L'utilisation des fichiers d'entrée (fichier de plaque) juste en début de traitement
- La mutualisation de flux sur un seul à l'aide d'une fonction de concaténation afin de simplifier le traitement
- La suppression des doublons au niveau des colonnes à envoyer par l'utilisation des fonctions de filtrage afin d'alléger le flux
- La réduction de la clé de filtrage sur 2 critères au lieu de 11
- L'envoi d'une seule requête vers la base de données au lieu de 5 pour chaque fichier de plaque
- La modification de la longueur des champs

2. Développement du DataMart Trafic ADSL

Je vais montrer en détail les différentes phases de réalisation d'un DataMart décisionnel avec l'ETL Talend Open Studio, de la modélisation de la chaîne de traitement à la construction du fichier du PARC_Client.txt.

2.1 Description du traitement

Ce projet consiste à construire une chaîne de traitement des données du trafic ADSL des clients d'Orange. Il est composé de deux jobs, chacun traitant une partie du traitement dont le résultat d'exécution sera utilisé comme entrée dans l'étape suivante.

Afin de générer le fichier PARC_Client.txt, il faut réaliser les opérations suivantes :

- Concaténer les cinq fichiers de plaque
- Extraire la liste des clients ayant une offre ADSL de type DATA à partir du flux venant des plaques
- Construire une clé de filtrage afin d'effacer les doublons
- Faire une requête vers la base de données Oracle contenant les types d'offre
- Etablir une jointure entre le flux principal et la base de données afin de compléter le flux par les informations sur les offres commerciales
- Dissocier les 4 types d'offres (DATA, TOIP, MLT, MLV)
- Construire la liste des services (concaténation des 4 offres)
- Générer le fichier PARC_Client.txt

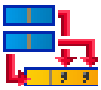
2.2 ETL Talend

Le choix d'une solution logicielle gratuite, qui offre les fonctionnalités nécessaires pour répondre aux besoins d'entreposage, fait de Talend open studio une solution adaptée aux besoins.

Etant une solution open source, Talend présente plusieurs avantages par rapport aux autres ETL :

- Support de nombreux systèmes : plus de 400 connecteurs
- Support, contributions : la force d'une communauté

2.3 Composants Talend utilisés

 **tFileInputDelimited**	Lit un fichier délimité ou un flux de données ligne par ligne, afin de le diviser en champs et d'envoyer ses champs au composant suivant, comme défini par le schéma, via une connexion Row.
 **tFileOutputDelimited**	Ecrit un fichier délimité contenant des données organisées en fonction du schéma défini.
 **tUnit**	Centralise des données provenant de sources diverses et hétérogènes (concaténation de flux structurés de manière identique).
 **tUniqRow**	Compare les entrées et supprime les doublons du flux d'entrée en fonction d'une clé.
 **tMap**	Transforme et dirige les données à partir d'une ou plusieurs source(s) et vers une ou plusieurs destination(s).
 **tDenormalize**	Dénormalise un flux entrant en fonction d'une colonne.
 **tSortRow**	Trie les données d'entrée basées sur une ou plusieurs colonnes, selon un type de tri (inner, outer).
 **tMysqlInput**	Lit une base de données MySQL et en extrait des champs à l'aide de requêtes SQL.

3. Scénario de mise en œuvre

Le processus d'extraction, de transformation et de chargement des données a été réalisé avec l'outil open source Talend Open Studio. Lors des traitements, l'outil Talend est capable de fonctionner avec un référentiel qui permet de centraliser les transformations et les tâches réalisées.

Plusieurs transformations ont été mises en place, en vue :

- d'extraire les données depuis des fichiers textes ou d'une base de données Oracle en spécifiant les champs désirés
- de transformer certains champs
- de filtrer selon différentes conditions
- de trier
- de joindre les données entre elles, ce qui facilite le chargement des tables du DataMart
- de sélectionner et/ou renommer des champs
- d'insérer et/ou de mettre à jour des tables…

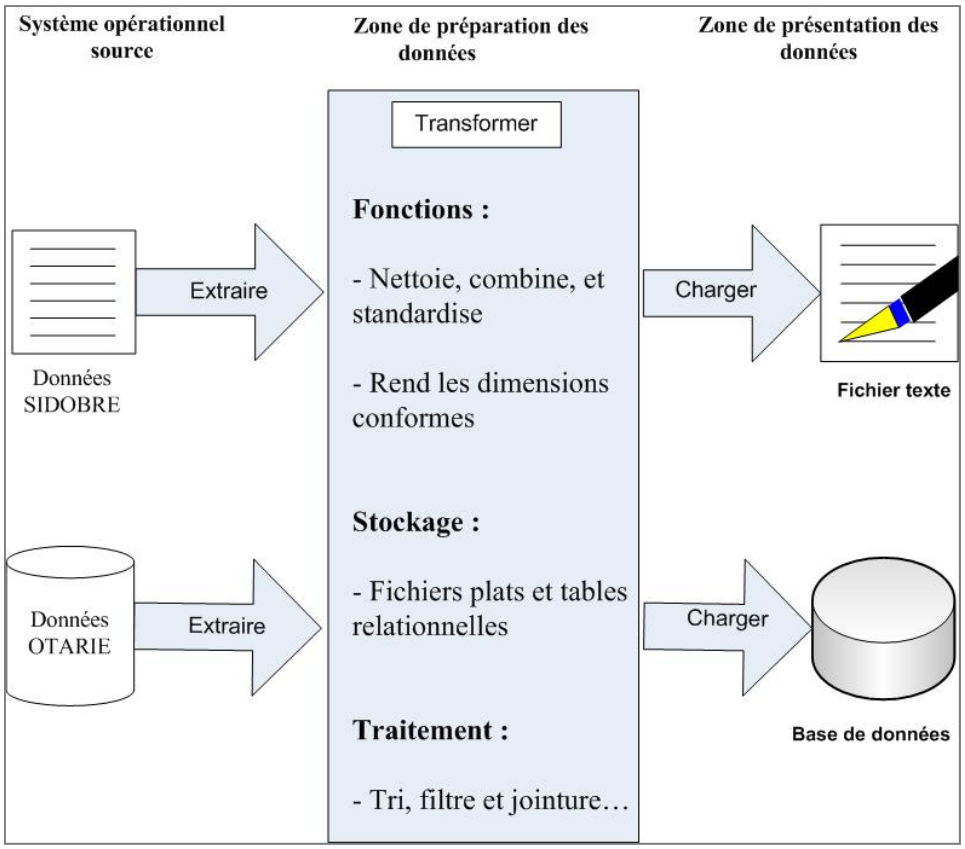


Figure 18 : COMPOSANT D'UN DATAMART TRAFIC ADSL

3.1 Extraction

Cette étape consiste à collecter les données nécessaires à l'alimentation du DataMart et à fournir des rapports.

3.1.1 Données source de traitement

Fichiers de plaques :

Le territoire national a été découpé par Orange en 5 plaques représentant les 5 régions géographique :

- U1 : plaque Ile-de-France
- OU : plaque Ouest
- NE : plaque Nord-Est
- SE : plaque Sud-Est
- SO : plaque Sud-Ouest

Les fichiers de plaques contiennent les informations techniques des clients (numéro de téléphone) pour un jour donné.

NOM	DESCRIPTION	TYPE
PLAQUE	Code de la plaque	VarChar(3)
DR	Code **DR (Direction Régionale)**	VarChar(2)
NRA_HD	Code **NRAHD (Code Boucle locale)**	VarChar(3)
NOMNRA	Nom du **NRA (Nom de la boucle locale)**	VarChar
DSLAMN	Nom du **DSLAMN**	VarChar(8)
DSLAM	Nom du **DSLAM**	VarChar(8)
VP_VLAN	Code du **VP_VLAN**	VarChar
OFFRE	Type de l'offre (DATA, TOIP, MLT, MLV)	VarChar(32)
EPC	Valeur contenant le **ND**	VarChar
ND	**N**uméro de **D**ésignation (numéro de téléphone)	VarChar(9)
NOMBAS	Nom du **BAS**	VarChar(8)
CHASSIS	Numéro du châssis	VarChar
CARTENUMAS	Numéro de la carte dans le châssis	VarChar(3)
PORNUMBAS	Numéro du port	VarChar(3)

VP	Code **VP (canaux virtuel ATM)**	VarChar(4)
VC	Code **VC (canaux virtuel ATM)**	VarChar(6)
FAI	Nom du fournisseur d'accès internet	VarChar(32)
REVENTE	Contient la valeur **NULL** ou **1**	Integer
DATEMES	Date de mise en service	VarChar(8)
ICC	**I**dentifiant	VarChar(32)

3.1.2 La Base de Données OTARIE

Afin de tester l'ETL Talend, d'étudier la faisabilité du portage du projet sous Talend et faciliter la présentation du travail réalisé sous forme de prototype, j'ai utilisé une base de données MySQL qui reprend la même structure que celle de la base de données Oracle utilisée par Orange Labs.

Pour la création de ma base de données MySQL, j'ai utilisé l'outil WAMP qui permet la création de bases de données avec PHP. J'ai donc créé la table nécessaire au fonctionnement de la chaîne de traitement afin de ne pas interférer sur la base de production.

Cette table contient les informations de type d'offres des clients dans le référentiel OTARIE.

NOM	DESCRIPTION	TYPE	Nullable
ID_OFFRE	N° de l'offre	NUMBER(4)	N
LIB_OFFRE	Nom de l'offre	VARCHAR2(32)	N
ID_OFFRE_OTARIE	N° de l'offre correspondant dans la table OFFRE d'OTARIE (s'il existe)	NUMBER(4)	O
LIB_OFFRE_OTARIE	Nom de l'offre correspondante dans la table OFFRE (s'il existe)	VARCHAR2(32)	O
DEBIT	Débit de l'offre	VARCHAR2(32)	O

3.2 Traitement

Cette phase intermédiaire dans le fonctionnement d'un ETL s'avère importante et demande beaucoup de réflexion et de précision. L'exécution de chaque étape doit être vérifiée car les données chargées dans le Datamart doivent évidemment être complètes et exactes.

Pour des raisons d'insuffisance de mémoire, j'ai dû scinder le traitement en deux jobs pour obtenir à la fin le fichier PARC_Client.txt contenant les informations sur les clients ainsi que leur liste de services associés.

3.2.1 JOB_1 : Génération du Flux_Principal

Ce job consiste à générer un flux correspondant au parc national des clients ADSL et de compléter ce flux avec les informations de type offre (DATA, TOIP, MLT, MLV) provenant de la base de données.

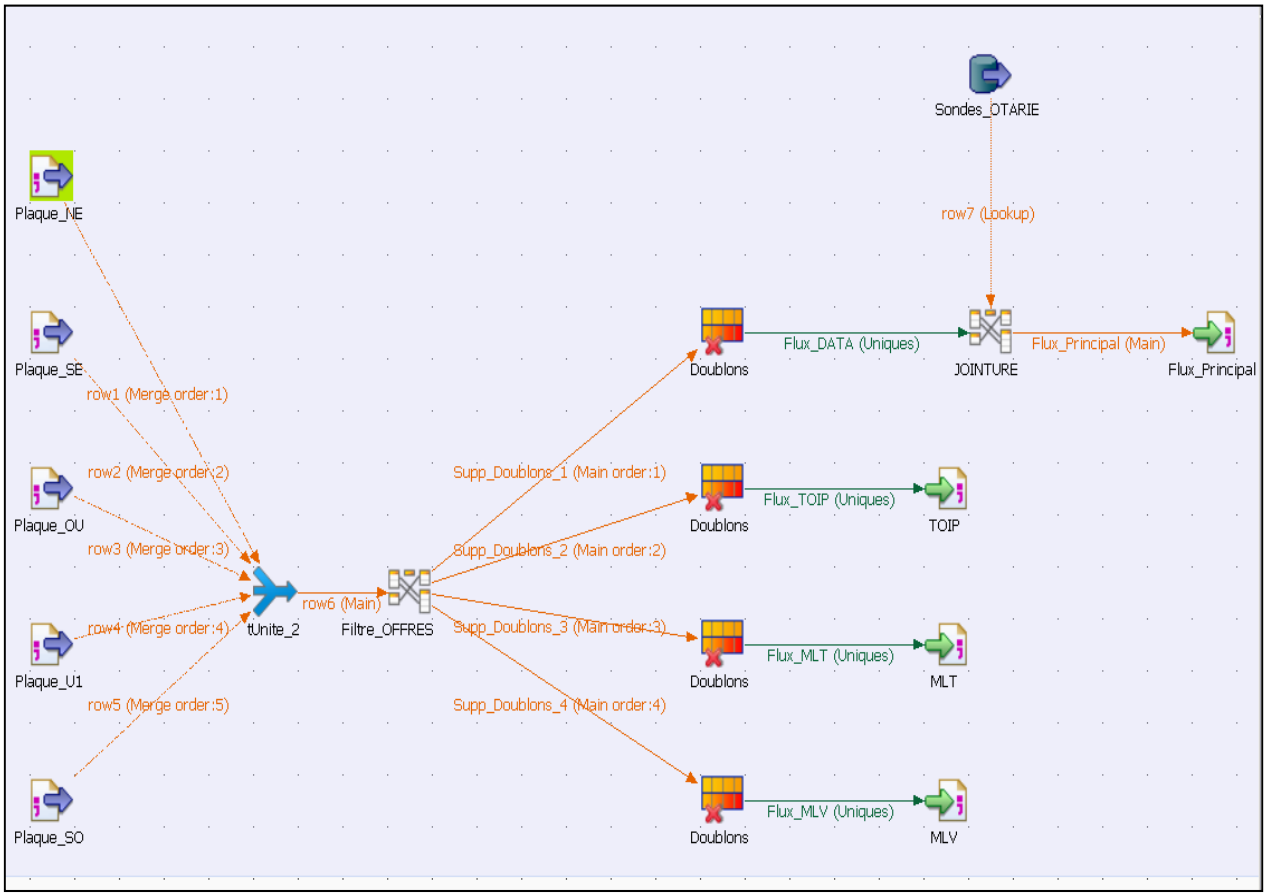


Figure 19 : GENERATION DU FLUX PRINCIPAL

Pour des besoins de jointure, de filtre et de tri, Talend dispose de composants simples. En revanche, il existe des composants multifonctions avec plusieurs entrée/sortie. C'est le cas du composant tMap qui, à lui seul, peut implémenter des fonctions de filtrage, de jointure et de tri.

3.2.1.1 Les éléments de traitement

En entrée de ce job, on utilise 5 fichiers de plaques de structure identique. Le traitement s'effectue comme suit :

- Lecture des 5 fichiers de plaque indépendamment, puis concaténation des flux à l'aide du composant tUnit, qui permet le réassemblage sur un seul flux et de n'en conserver que les données utiles.

- Tous les clients ont une offre DATA. Pour ne garder qu'une ligne par client, on filtre sur les offres (DATA, TOIP, MLT, MLV) à l'aide du composant tMap à partir du flux venant du composant tUnit.

  Chaque offre est reconnaissable par son libellé. Le filtrage effectué au niveau du composant tMap Filtre_OFFRES génère quatre flux correspondant aux 4 types d'offre possibles (MLV, MLT, TOIP, DATA).

- Un fichier de plaque peut contenir plusieurs lignes par client. Afin d'éviter les doublons, j'ai utilisé une clé de filtrage représentative avec moins de critères que dans le traitement actuel (2 critères au lieu de 11). J'ai également constaté que Talend gère mal la multitude de critères. Par conséquent, le choix de cette clé de filtrage était approprié.

  La suppression des doublons se fait à l'aide du composant tUniqRow sur les critères suivants :

  - ✓ **ND**
  - ✓ **DATEMES**

- En sortie de ce composant, les flux TOIP, MLV et MLT seront envoyés dans des composants tFileOutputDelimited afin de générer des fichiers au format CSV pour les utiliser comme entrée du JOB_2.

- Le flux DATA est envoyé vers un composant tMap afin de l'enrichir des données de l'application OTARIE. Celles-ci viennent du composant tMysqlInput qui permet de compléter le flux DATA par les informations de type OFFRE issues de l'application OTARIE. La requête est la suivante :

  Select **LIB_OFFRE, LIB_OFFRE_OTARIE** From **otarie**

- Ensuite, une jointure est faite sur la colonne OFFRE venant du Flux_Principal et la colonne LIB_OFFRE venant de la base de données à l'aide du composant tMap, afin d'ajouter le libellé de l'offre OTARIE au Flux Principal.

- En sortie du tMap, le Flux_Principal est envoyé sur un composant tFileOutputDelimited pour générer également un fichier CSV qui servira d'entrée principale dans le job suivant.

3.2.2 JOB_2 : Construction de la liste des services

Le job est composé des résultats d'exécution du JOB_1, avec en entrée le fichier contenant les données du Flux_Principal ainsi que les fichiers TOIP, MLT et MLV.

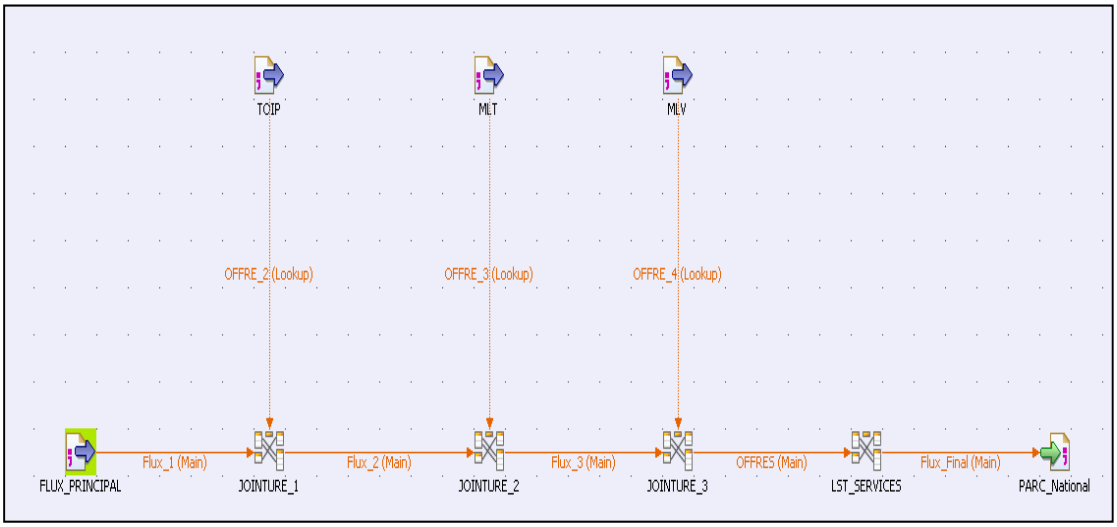


Figure 20 : GENERATION DU FICHIER DE PARC CLIENT

3.2.2.1 Les éléments de traitement

- En entrée de ce job, j'utilise le composant tFileInputDelimited pour lire le fichier Flux_Principal qui contient des informations sur le client ainsi que son offre DATA présenté par la colonne OFFRE_1.

- J'utilise aussi le composant tFileInputDelimited du flux TOIP qui contient la liste des clients ayant une offre TOIP.

- J'établis une jointure sur le ND entre les 2 flux venant de FLUX_PRINCIPAL et TOIP, en utilisant le composant tMap JOINTURE_1 afin de compléter le flux principal par la colonne OFFRE_2 contenant les offres TOIP de chaque client.

- Le même traitement est effectué pour les OFFRE_3 contenant les offres MLT et OFFRE4 contenant les offres MLV dans les composants tMap correspondants (JOINTURE_2 et JOINTURE_3).

- Pour éviter d'avoir des colonnes nulles, j'ai mis pour chaque offre la condition suivante : « Si on trouve des valeurs nulles, on les remplace par une chaîne vide, sinon on garde le contenu de la colonne ». Cela se traduit par :

Relational.ISNULL(row1.monChamp)?"":" row1.monChamp "

A la fin de ce traitement, on obtient un flux contenant toutes les informations sur les clients ainsi que les 4 offres auxquelles ils ont souscrit. Ensuite, on passe à la construction de la liste des services.

- Par concaténation des champs OFFRE_1, OFFRE_2, OFFRE_3 et OFFRE_4, on constitue le champ LST_SERVICES en utilisant la syntaxe suivante :

"#"+OFFRES.OFFRE_1+"#"+OFFRES.OFFRE_2+"#"+OFFRES.OFFRE_3+"#" +OFFRES.OFFRE_4+"#

A la fin de ce job, on alimente le fichier de PARC_Client.txt avec la liste des services de chaque client.

3.3 Chargement

L'objectif principal de cette étape d'alimentation du fichier PARC_Client.txt est de rassembler les données collectées d'une manière cohérente, simple à utiliser pour les mettre à disposition des utilisateurs internes du groupe Orange.

Les informations contenues dans le fichier de PARC_Client.txt sont les suivantes :

NOM	DESCRIPTION	TYPE
ND	**N**uméro de **D**ésignation	VarChar(9)
VP	Code **VP**	VarChar(4)
VC	Code **VC**	VarChar(6)
NOMBAS	Nom du **BAS**	VarChar(8)
CARTENUMBAS	Numéro de la carte dans le châssis	VarChar(3)
PORTNUMBAS	Numéro du port	VarChar(3)
LST_SERVICES	Liste des services	VarChar(128)
LIB_OFFRE_OTARIE	Libellé de l'offre **OTARIE**	VarChar(32)
DATEMES	Date de mise en service	VarChar(8)
DSLAMN	Nom du **DSLAMN**	VarChar(8)
DSLAM	Nom du **DSLAM**	VarChar(8)
ICC	**I**dentifiant	VarChar(32)

Difficultés techniques

Les difficultés techniques sont liées à toute reprise d'une chaîne existante. Le modèle de développement, les contraintes du schéma de la base de données ainsi que l'ETL utilisé sont les principaux points que l'on peut citer :

- L'apprentissage des outils ETL DataStage et Talend Open Studio, totalement inconnus de moi, a été une des premières difficultés. J'ai dû suivre une auto-formation, aidé par le support de cours, avant de pouvoir m'attaquer à la problématique du sujet de stage.

- La mise en œuvre de la chaîne avec un nouvel outil s'est retrouvée compliquée par la recherche des composants adéquats pour réaliser les traitements nécessaires.

- L'intervention fréquente sur la base de données pour la mettre à jour m'a permis de me familiariser avec la structure des données et de mieux comprendre le fonctionnement de la chaîne de traitement.

- L'accomplissement des tâches et la réalisation des objectifs fixés par les clients internes d'Orange ont parfois provoqué des complications (format du champ Lst_Services)

- L'espace mémoire n'étant pas toujours suffisant sur ma machine, j'ai dû mettre en œuvre les solutions techniques suivantes :

  - ✓ Utilisation de l'espace disque (paramètre du composant tMap)
  - ✓ Découpage des jobs

  Une autre solution aurait été d'utiliser une machine plus puissante et avec une capacité de mémoire plus importante.

Bilan Personnel

Ce stage a été un excellent complément à ma formation de Master. Il m'a permis d'affronter les connaissances et les méthodes de travail que j'avais acquises tout au long de mes études, avec la réalité des entreprises. En effet, pendant ce stage, j'ai pu développer en particulier des compétences techniques et relationnelles très importantes pour mon futur professionnel.

J'ai amélioré aussi ma capacité d'écoute pour savoir en déduire les besoins des clients et ma capacité d'analyse et de synthèse. J'ai dû travailler sur mon expression orale afin de savoir m'exprimer de façon claire et directe et j'ai appris à savoir prioriser les différents moyens de communication (oraux et écrits) qui existent dans l'entreprise.

Pendant ce stage, j'ai travaillé de façon autonome et en équipe. Quant au travail en équipe, je trouve qu'il est très enrichissant puisque l'on apprend beaucoup des échanges que l'on a avec les autres membres de l'équipe. J'ai beaucoup apprécié de pouvoir travailler avec différentes personnes car chacune d'elles a une façon unique de travailler et d'affronter les problèmes rencontrés.

J'ai appris à prioriser les tâches, à bien organiser mon temps tout en restant flexible et à m'adapter aux rythmes de travail et aux exigences des personnes avec qui je travaillais.

Je me suis sentie très bien accueillie par les personnes du secteur car il y a une ambiance de travail à la fois très professionnelle et très humaine. Je me suis rendu compte de l'importance d'être à l'aise dans son environnement de travail et d'apprécier la compagnie des collègues en dehors des contextes strictement professionnels. Ce stage m'a permis de m'épanouir aussi bien sur le plan professionnel que personnel.

Conclusion

Les nouvelles technologies de l'information permettent de concevoir des systèmes d'informations particulièrement performants. Ces derniers fournissent d'importantes informations mais ne sont pas conçus pour permettre leur utilisation dans un processus d'aide à la décision.

Aussi, le DataMart permet au décideur de travailler dans un environnement informationnel, référencé, homogène et historisé. Cette technique l'affranchit des problèmes liés à l'hétérogénéité des systèmes informatiques, ainsi que celle des différentes données issues de l'organisation.

Ainsi, confronté à un environnement de plus en plus concurrentiel, le groupe Orange s'est doté d'outils performants et a mis en place un DataMart robuste. Mes travaux m'ont permis de faire une étude de faisabilité sur un DataMart existant, puis l'analyse, la conception et le développement d'une solution sur une nouvelle plateforme. Cette étude était indispensable afin de cibler les améliorations possibles, de rationaliser le traitement et d'étudier l'utilisation d'un ETL open source en remplacement de la solution propriétaire actuellement mise en œuvre sur le DataMart Trafic ADSL des clients d'Orange.

Ce projet m'a permis de faire évoluer considérablement mes connaissances et mes compétences dans le domaine de l'informatique décisionnelle et des outils associés (base de données, ETL…). J'ai approfondi mes compétences en configuration, en intégration de bases de données et, sur la partie modélisation Business Intelligence, j'ai également amélioré mes capacités en gestion de projet. J'ai pu développer mon autonomie et mon esprit d'analyse. Ce stage me permet d'être préparée à faire face aux besoins des entreprises dans le domaine de la BI.

Abréviations et acronymes

Termes	Libellé
ADSL	**A**symetric **D**igital **S**ubscriber **L**ine
ATM	**A**synchronous **T**ransfer **M**ode
BAS	**B**roadband **A**ccess **S**erver
BI	**B**usiness **I**ntelligence
DATEMES	**D**ate **M**ise en **S**ervice
DSLAM	**D**igital **S**ubscriber **L**ine **A**ccess **M**ultiplexer
ERP	**E**nterprise **R**esource **P**lanning
ETL	**E**xtraction, **T**ransformation & **C**hargement
FAI	**F**ournisseur d'**A**ccès **I**nternet
GE	**G**igabits **E**thernet
MLT	**M**a **L**igne **T**V
MLV	**M**a **L**igne **V**isio
MOA	**M**aitre d'**O**uvrage
ND	**N**uméro de **D**ésignation
SAS	**S**tatistical **A**nalysis **S**ystem
SGBD	**S**ystème de **G**estion de **B**ase de **D**onnées
SSII	**S**ociété de **S**ervices en **I**ngénierie **I**nformatique
TOIP	**T**éléphonie sur **IP**
VC	**V**irtual **C**hannel
VoIP	**Vo**ix sur **IP**
VP	**V**irtual **P**ath
WAMP	**W**indows, **A**pache, **M**ySQL, **P**HP

Glossaire

ADSL
Technologie appartenant à la famille désignée sous le nom générique de xDSL, qui regroupe un ensemble de systèmes destinés à augmenter les performances du réseau téléphonique existant.

Alimentation
Insertion de données d'une source vers une cible. Dans un projet décisionnel, on observe en général plusieurs bases de données source et une base cible (Datamart). Ces informations peuvent subir des transformations avant leur chargement.

Base de données
Ensemble de données structuré, enregistré sur un système de fichier. Ces informations sont gérées à l'aide d'un Système de Gestion de Base de Données.

Business Intelligence
Dénomination anglophone de « décisionnel ». C'est un domaine dont l'objectif est d'obtenir une vue d'ensemble d'une activité, par le biais de rapports, tableaux de bord… Cela consiste en la collection, la consolidation, la modélisation et la restitution des données.

Composant
Connecteur pré-configuré exécutant une opération spécifique d'intégration de données, quel que soit le type de données (bases de données, applications, fichiers plats, services Web, etc.).

Datamart
Aussi appelé « Magasin de données », le Datamart est un ensemble de données ciblées (qui concernent un métier ou domaine spécifique), organisées, regroupées et agrégées (catégorisées).

Elément
Unité technique constituant un projet. Les éléments sont regroupés en fonction de leur type : Job Design, Business Model, Context, Code, Metadata, etc.

ETL
Permet l'Extraction, la Transformation et le Chargement de données depuis des sources diverses (bases de données, fichiers) vers des cibles.

Intégrité des données
Préserver l'intégrité des données, c'est s'assurer que celles-ci ne subissent aucune altération et conservent un format permettant leur utilisation (pas de saisie de chaîne de caractères dans un champ numérique par exemple).

Job
Concept graphique, composé d'un ou plusieurs composants reliés entre eux. Il permet de mettre en place des processus opérationnels de gestion des flux.

MySQL
SGBD relativement puissant et léger utilisé principalement pour les sites internet, du fait de sa capacité à traiter de grands volumes de données. La version 5 de MySQL supporte les procédures stockées.

OTARIE
Application développée par Orange Labs pour collecter les données de trafic issues des sondes positionnées sur le réseau.

Projet
Ensemble structuré d'éléments techniques et de leurs métadonnées associées.

PHP
Langage de programmation qui embarque des fonctionnalités objet depuis sa version 5. Principalement utilisé pour les sites internet, il est également répandu parmi les applications web d'entreprise.

Rapport
Document édité par un logiciel. C'est une présentation périodique des bilans analytiques sur les activités et résultats d'une organisation. Ces données sont généralement issues d'un Datamart, d'un Datawarehouse ou d'une base de données en vue de fournir des résultats d'analyse.

Référentiel
Espace de stockage (repository en anglais) utilisé par Talend pour regrouper les données liées aux éléments techniques utilisés soit pour décrire les Business Models, soit pour créer les jobs.

Schéma
Définit le nom des champs, leur type et leur taille qui sont utilisés par un composant. Le schéma est soit local au composant (Built-in), soit commun à plusieurs composants (dans le Repository).

Séquenceur
Ordonnanceur découpé en une suite ordonnée d'opérations ou d'éléments pour le lancement des jobs à partir d'un script.

Stage
Elément d'un **job** sous DataStage (appelé également composant sous Talend Open Studio).

Système de Gestion de Base de Données (SGBD)
Logiciel permettant la manipulation des bases de données (consultation, modification, mise à jour, insertion, suppression…).

Bibliographie

Sites Internet :

Site officiel d'Orange :

- http://www.orange.com/fr_FR/groupe/

Site officiel Talend :

- http://www.talend.com
- http://www.talendforge.org

Wikipedia pour de nombreuses informations en tout genre :

- http://fr.wikipedia.org/wiki/Entreposage
- http://fr.wikipedia.org/wiki/Datamart

Forums et tutoriels :

- http://www.developpez.net/forums/d1050064/logiciels/solutions-dentreprise/business-intelligence/etl/talend/alimentation-datamart/
- http://business-intelligence.developpez.com/faq/talend/?page=II
- http://www.labdecisionnel.com

Des informations sur le décisionnel :

- http://www.decideo.fr/
- http://www.informatiquedecisionnelle.com/

Ouvrage :

"**Le Datawarehouse, le Data Mining**" - Jean-Michel Franco et EDS-Institut Prométhéus - Eyrolles, 1996

"**Datawarehouse et Data Mining**" - Conservatoire National des Arts et Métiers Juin 1998

"**La construction du datawarehouse**" - Jean-François Goglin – 2001

Résumé

Afin de valider un Master 2 Informatique en Systèmes Distribués et Réseaux à l'Université de Franche Comté, j'ai réalisé un projet décisionnel au sein d'Orange Labs à Belfort, l'un des centres de Recherche et Développement d'Orange.

Le sujet proposé consistait dans un premier temps à réaliser une étude de l'infocentre existant et d'effectuer un reverse engineering sur une chaîne de traitement. Cela m'a permis de m'imprégner du sujet, de prendre en main l'ETL DataStage et de produire un document décrivant la chaîne de traitement.

Le travail accompli dans la première partie a porté sur l'ETL DataStage utilisé par Orange Labs pour gérer cet infocentre. Il a consisté, après le recueil et l'analyse des besoins client, à cibler les axes d'amélioration possibles pour présenter la chaîne de traitement d'une manière plus structurée et facile à gérer. La conclusion de ce travail a été la rédaction d'un document analytique complet qui sera utilisé dans la suite du stage. La structure de la chaîne du traitement a été modifiée pour prendre en compte les améliorations et faciliter ensuite le travail de la phase suivante de conception du DataMart trafic.

Dans un second temps, il fallait concevoir un DataMart Trafic Clients ADSL centralisant les informations relatives aux activités des clients d'Orange à l'aide d'un outil open source.

Le travail consistait en particulier à faire l'analyse et la modélisation du DataMart, mais aussi le développement afin d'alimenter ce dernier. L'objectif était de faire une étude de faisabilité pour porter la chaîne sur un environnement open source, en l'occurrence Talend Open Studio.

Mots clefs : Décisionnel, DataMart, alimentation,

Technologies: MySQL V5, ORACLE, Data Stage, Talend Open Studio

Orange Labs Recherche et Développement
1 rue Maurice et Louis de Broglie
CS 20382

90007 BELFORT Cedex