



Mise en place d'une solution ETL dans le cadre de la rationalisation d'un système d'information

Frédéric Cesbron

► To cite this version:

Frédéric Cesbron. Mise en place d'une solution ETL dans le cadre de la rationalisation d'un système d'information. Génie logiciel [cs.SE]. 2015. dumas-01681770

HAL Id: dumas-01681770

<https://dumas.ccsd.cnrs.fr/dumas-01681770>

Submitted on 11 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

PARIS

MEMOIRE

présenté en vue d'obtenir

le DIPLOME D'INGENIEUR CNAM

SPECIALITE : Informatique

OPTION : Système d'information

par

Frédéric CESBRON

**Mise en place d'une solution ETL dans le cadre de la
rationalisation d'un système d'information**

Soutenu le 25 mars 2015

JURY

PRESIDENT :	Mme Isabelle WATTIAU	Professeur - CNAM Paris
MEMBRES :	M. Jacky AKOKA	Professeur - CNAM Paris
	Mme Faten ATIGUI	Maître de conférences - CNAM Paris
	Mme Marie-France BARAS	DSI - CGA
	M. Sébastien FOSSE	Chef de Projet Rationalisation - CGA

Remerciements

Ce mémoire est la conclusion d'un cursus commencé au CNAM Nantes en 2001. Durant ce long parcours, j'ai pu compter sur de nombreux soutiens.

Je remercie particulièrement l'ensemble des professeurs, intervenants et auditeurs que j'ai pu rencontrer et avec qui j'ai pu échanger pendant ces nombreuses années au CNAM Pays de la Loire et au CNAM Paris.

Je remercie également mes parents qui m'ont encouragé en 2001 à me lancer dans cette aventure CNAM.

Concernant le projet qui fait l'objet de ce mémoire, je tiens à remercier spécifiquement M. Akoka, professeur au CNAM Paris, qui m'a apporté conseils et suivi tout au long du projet.

Je souhaite également remercier tous les membres de la DSI de CGA qui savent maintenir une bonne ambiance au sein du service qui, de fait, est propice à un travail efficace.

Je remercie Marie-France Baras, Directrice des Systèmes d'Information de CGA, qui aura été mon sponsor sur ce projet et qui aura su être compréhensive de par son expérience d'auditrice CNAM.

Enfin, un grand merci à ma compagne qui a su entretenir mon niveau de motivation et d'investissement pendant toute cette année de travail. Elle m'a soutenu tout au long de ce projet malgré les impacts sur nos loisirs, vacances...

Liste des abréviations

AD :	Active Directory
API :	Application Programming Interface
BDDF :	Banque de Détail en France
CFONB :	Comité Français d'Organisation et de Normalisation Bancaires
CFT :	Cross File Transfer
CGA :	Compagnie Générale d'Affecturage
CSV :	Comma-Separated Values
DAT :	Dossier d'Architecture Technique
DBA :	DataBase Administrator
DCOM :	Distributed Component Object Model
DICP :	Disponibilité Intégrité Confidentialité Preuve
DNS :	Domain Name System
DPR :	Direction des PRogrammes
DRC :	Direction de la Relation Client
DRSP :	Direction des Règlements et Services Partagés
DSI :	Direction des Systèmes d'Information
EAI :	Enterprise Application Integration
EEI :	Enterprise Information Integration
EJB :	Enterprise JavaBeans
ETL :	Extract Transform and Load
ETP :	Equivalent Temps Plein
FCI :	Factors Chain International
FTP :	File Transfer Protocol
GBIS :	Global Banking and Investor Solutions

GED :	Gestion Electronique des Documents
GTB :	Global Transaction Banking
GTPS :	Global Transaction and Payment Services
HTTP :	HyperText Transfer Protocol
IBFS :	International Banking and Financial Services
IFS :	Integrated File System
JDBC :	Java DataBase Connectivity
JEE :	Java Enterprise Edition
JMS :	Java Message Service
JPEG :	Joint Photographic Experts Group
LDAP :	Lightweight Directory Access Protocol
MOA :	Maitrise d'Ouvrage
MOE :	Maitrise d'œuvre
MOM :	Message Oriented Middleware
ODBC :	Open DataBase Connectivity
OSI :	Opérations Structurées et Internationales
PDF :	Portable Document Format
PMBOK :	Project Management Body Of Knowledge
POC :	Proof of Concept
PRA :	Plan de Reprise d'Activité
ROI :	Return On Investment
RPG :	Report Program Generator
SEPA :	Single Euro Payments Area
SFTP :	Secure File Transfer Protocol
SG :	Société Générale
SGBD :	Système de Gestion des Bases de Données

SI :	Système d'information
SQL :	Structured Query Language
SSO :	Single Sign On
SWIFT :	Society for Worldwide Interbank Financial Telecommunication
TIFF :	Tagged Image File Format
VM :	Virtual Machine
XML :	eXtensible Markup Language

Table des matières

Remerciements	2
Liste des abréviations	3
Table des matières	6
Introduction	8
I LE CONTEXTE	9
I.1 LA SOCIETE GENERALE.....	9
I.2 LA COMPAGNIE GENERALE D’AFFACTURAGE.....	10
I.2.1 Les métiers de CGA.....	10
I.2.2 Focus sur l’affacturage.....	11
I.2.3 L’organisation.....	12
I.2.4 La direction des systèmes d’information	14
I.2.5 Le système d’information	16
I.2.6 Le schéma directeur	18
II LA PROBLEMATIQUE	20
II.1 LES ENJEUX	20
II.2 FOCUS SUR LES SOLUTIONS D’INTEGRATION DE DONNEES	22
II.2.1 La définition des ETL	22
II.2.2 Les normes et standards	22
II.2.3 Les tendances	23
II.2.4 Les outils et acteurs.....	24
III L’ACQUISITION D’UNE SOLUTION ETL	26
III.1 L’APPEL D’OFFRE	26
III.2 LA REPONSE AU BESOIN ET LE CHOIX DE LA SOLUTION	36
III.3 LA PRESENTATION GENERALE DE LA SOLUTION RETENUE	39
III.4 LE BILAN DE LA PHASE D’ACQUISITION	45
IV LA MISE EN PLACE DE LA SOLUTION	46
IV.1 L’INSTALLATION DE LA SOLUTION CHEZ CGA	46
IV.2 L’INTEGRATION DE L’ETL DANS LE SI DE CGA.....	50
IV.2.1 La gestion des droits	50
IV.2.2 La mise en production des jobs	51
IV.2.3 L’exécution des jobs	52
IV.2.4 La supervision des jobs	52
IV.2.5 Le kit méthodologique	53
IV.3 LA MISE EN PLACE DES BONNES PRATIQUES D’UTILISATION DE LA SOLUTION	55
IV.3.1 La gestion des projets.....	55
IV.3.2 La conception des travaux Datastage	55
IV.3.3 L’exploitation de la solution	57
IV.4 LE BILAN DE LA PHASE DE MISE EN PLACE.....	58
V LA REALISATION DES PREMIERS PARAMETRAGES	59
V.1 LE LOT 1 : PROJETS CONTRAINTS.....	59
V.1.1 La migration d’un client majeur sur une nouvelle plateforme de financement.....	59
V.1.2 L’intégration d’une nouvelle filiale	67
V.1.3 La refonte complète du site transactionnel de CGA	70
V.2 LE LOT 2 : RATIONALISATION / ACCOMPAGNEMENT	74

V.2.1	Le remplacement d'un outil ETL existant	74
V.2.2	L'étude pratique sur les fichiers multi-format	80
V.2.3	L'étude sur l'intégration des fichiers volumineux	83
V.2.4	La mise en place de templates	87
V.3	LE BILAN DE LA PHASE REALISATION	90
VI	LE PILOTAGE DU PROJET	91
VI.1	LE LANCEMENT DU PROJET.....	91
VI.2	LE SUIVI DU PROJET.....	94
VI.3	LA CLOTURE DU PROJET	96
VII	LE BILAN	98
VII.1	LE BILAN DU PROJET.....	98
VII.2	LE BILAN PERSONNEL	101
VII.3	LES PERSPECTIVES.....	102
Conclusion.....		103
Références		104
Table des annexes.....		105
Annexe 1 Adéquation aux besoins		106
Annexe 2 Analyse des coûts.....		108
Annexe 3 Classement des réponses		109
Annexe 4 Business Case.....		111
Annexe 5 Format du fichier Monoprix en entrée		113
Annexe 6 Format du fichier Monoprix en sortie		114
Annexe 7 Format du fichier d'items multi-format		115
Annexe 8 Matrice des responsabilités		116
Annexe 9 Fiche de Suivi de Projet		118
Annexe 10 Reste A Faire.....		119
Liste des figures.....		120
Liste des tableaux		121

Introduction

La Compagnie Générale d’Affacturage aborde une nouvelle phase de son essor, passant par l’exploitation de nouveaux relais de croissance et l’amplification de son rôle au sein du groupe Société Générale.

Pour assurer l’atteinte de ses objectifs stratégiques, CGA doit adapter son système d’information notamment en termes d’agilité, de simplicité et d’automatisation. Dans ce contexte, l’établissement et l’acceptation d’un schéma directeur du système d’information est en cours depuis le début de l’année 2014.

L’un des aspects principaux de ce schéma directeur est l’acquisition d’outils performants et intégrés permettant de simplifier et de rationaliser durablement le système d’information.

Pour répondre à cet objectif, un ensemble de projets a été identifié parmi lesquels celui qui fait l’objet de ce mémoire et qui consiste à mettre en place une solution d’intégration des flux de données basée sur un outil ETL (Extract – Transform – Load) du marché.

Le projet qui m’a été confié consiste donc à choisir un outil ETL adapté au contexte de CGA, à le mettre en place, à l’intégrer au système d’information, à définir les bonnes pratiques et à réaliser les premiers développements avec cet outil.

Ce mémoire rend compte des travaux réalisés pour répondre à ces objectifs. Il est structuré en 7 parties. La première partie décrit le contexte dans lequel ce projet est mené. La seconde partie détaille la problématique à l’origine de ce projet et en décrit les enjeux. La troisième partie est consacrée à la phase d’acquisition de l’outil menant au choix de l’outil. La quatrième partie définit la mise en place de l’outil choisi au sein du système d’information de CGA. La cinquième partie présente les premiers développements réalisés grâce à la solution choisie. La sixième partie explique la méthode de pilotage du projet. Enfin, la dernière partie est consacrée au bilan du projet pour CGA mais également d’un point de vue personnel.

I Le contexte

I.1 La Société Générale

Société Générale est l'un des premiers groupes européens de services financiers. S'appuyant sur un modèle diversifié de banque universelle, le Groupe allie solidité financière et stratégie de croissance durable avec l'ambition d'être la banque relationnelle, référence sur les marchés, proche de ses clients, choisie pour la qualité et l'engagement de ses équipes.

Acteur de l'économie réelle depuis 150 ans, Société Générale emploie plus de 154 000 collaborateurs, présents dans 76 pays, et accompagne au quotidien 32 millions de clients dans le monde entier en offrant une large palette de conseils et solutions financières sur mesure aux particuliers, entreprises et investisseurs institutionnels, qui s'appuie sur trois pôles métiers complémentaires :

- BDDF (Banque de Détail en France) : La banque de détail en France avec les enseignes Société Générale, Crédit du Nord et Boursorama qui offrent des gammes complètes de services financiers avec une offre multi canal à la pointe de l'innovation digitale. Au sein de ce pôle, GTPS (Global Transaction & Payment Services) offre une gamme complète et intégrée de services reposant sur l'expertise des métiers de Transaction Banking et de Payment Services. GTB (Global Transaction Banking) intègre l'ensemble des expertises et des services destinés à faciliter en domestique et à l'international les activités transactionnelles quotidiennes des grands acteurs économiques et financiers.
- IBFS (International Banking and Financial Services) : La banque de détail à l'international, services financiers et assurances avec des réseaux présents dans les zones géographiques en développement et des métiers spécialisés leaders dans leurs marchés.
- GBIS (Global Banking and Investor Solutions) : La banque de financement et d'investissement, banque privée, gestion d'actifs et métier titres avec leurs expertises reconnues, positions internationales clés et solutions intégrées.

I.2 La Compagnie Générale d’Affacturage

La CGA (Compagnie Générale d’Affacturage) est une filiale à 100% du groupe Société Générale depuis 1974 d’abord sous la marque Soge-Factoring puis sous la dénomination CGA à partir de 1992. Elle est présente depuis 2005 sur l’île de la Réunion et en Allemagne depuis 2014. L’effectif de CGA est de 330 collaborateurs. CGA est rattachée à la branche GTB du pôle GTPS du groupe Société Générale. Elle évolue sur le marché français et à l’international et adhère au réseau de factors international FCI.

En termes de performance, CGA occupe la 3^{ème} position en volume de créances achetées sur le marché français de l’affacturage.

I.2.1 Les métiers de CGA

En tant que spécialiste de la sphère des métiers du crédit interentreprises, CGA propose à ses clients quatre principaux services en France et à l’international :

- l’optimisation financière au travers du financement de leur poste Clients, la mise en place de programmes à caractère déconsolidant
- la gestion administrative et comptable du poste Clients et son recouvrement
- la sécurisation de leur développement commercial par la couverture de leurs risques clients
- la mise en œuvre de programmes de financement du crédit fournisseur (Supply Chain Finance)

CGA dispose d’une des gammes de produits la plus large du marché de l’affacturage. L’offre de CGA est destinée à toute taille d’entreprises (de la PME à la multinationale) et aux associations. CGA totalise ainsi 4800 clients correspondants à des marchés très diversifiés.

I.2.2 Focus sur l'affacturage

L'affacturage peut être défini comme « une technique qui permet à un vendeur de biens ou services de transférer ses créances professionnelles sur ses acheteurs à un organisme spécialisé, le factor, qui les prend en charge » [02].

L'affacturage est donc une relation tripartite décrite par le schéma suivant.

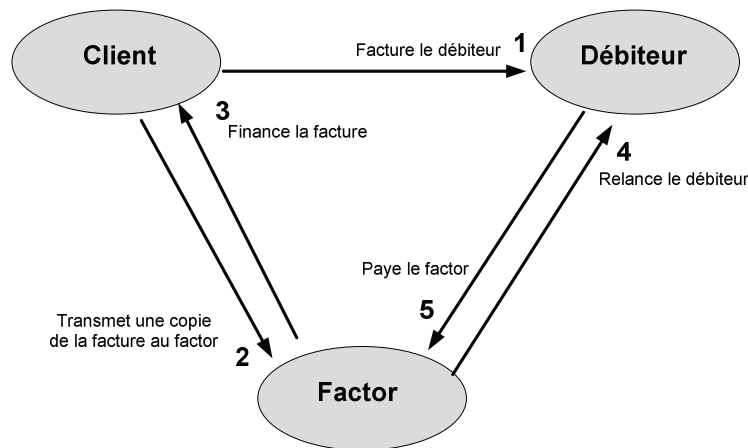


Figure 1 : Principes de l'affacturage

Le factor propose 3 services principaux à l'entreprise cliente : le financement des créances, la gestion du poste client et la garantie contre le risque d'impayés.

Cette définition simplifiée de l'affacturage est bien sûr compliquée par la réalité du marché qui implique des montages plus complexes en fonction des spécificités des adhérents.

On peut notamment citer :

- l'affacturage export : l'adhérent travaille avec un client (débiteur) localisé dans un pays étranger. Le factor n'ayant pas les moyens techniques de faire du recouvrement à l'étranger, un contact est pris avec un factor correspondant (factor du pays du débiteur)
- l'affacturage import : l'adhérent localisé à l'étranger travaille avec un client (débiteur) situé en France. Le factor étranger n'ayant pas les moyens techniques de faire du recouvrement en France, il prend contact avec le factor français.
- l'affacturage inversé : la relation d'affacturage est entre le débiteur (généralement un donneur d'ordre important) et le factor.
- l'affacturage non notifié : le client ne souhaite pas divulguer au débiteur son contrat d'affacturage. Dans ce cas, la gestion du poste client est conservée par le client.

I.2.3 L'organisation

L'organisation de CGA respecte l'organigramme ci-dessous.

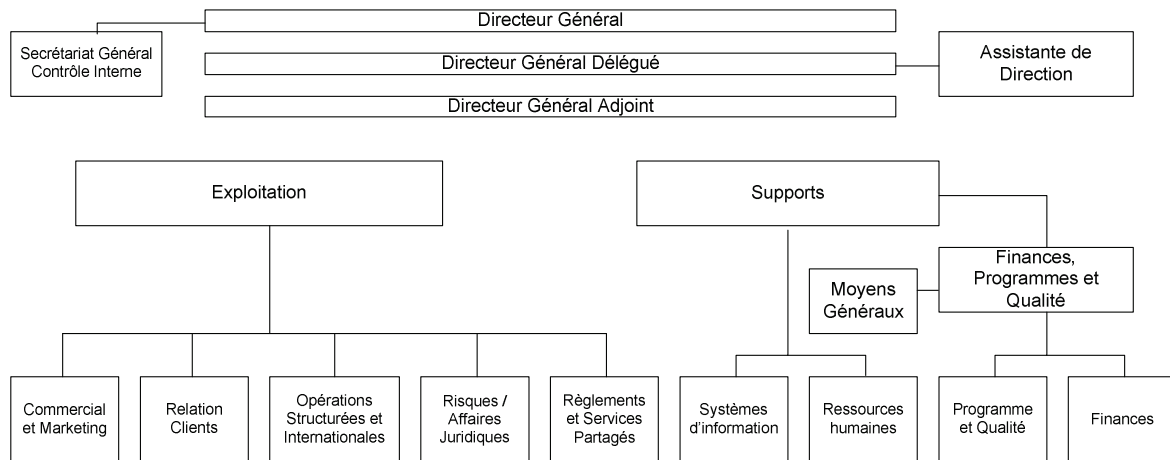


Figure 2 : Organigramme CGA

La direction de l'exploitation regroupe :

- **La direction commerciale**, déployée dans les délégations régionales du groupe Société Générale, assure la présence de CGA dans les agences du réseau au contact des prospects. Elle distribue aussi les produits CGA auprès d'entreprises non clientes du réseau SG au moyen d'une force de vente dédiée et implantée en région. Cette force de vente s'appuie sur un réseau d'apporteurs (Crédit du Nord, Crédit Mutuel Arkea, courtiers, etc...)
- **La direction relation clients (DRC)** assure la relation avec l'adhérent au quotidien grâce aux gestionnaires et chargés de clientèle ainsi que le recouvrement des créances. Elle supervise hiérarchiquement la succursale réunionnaise.
- **La direction des risques et affaires juridiques** garantit aux clients la préservation de leurs intérêts vis-à-vis des débiteurs.
- **La direction opérations structurées et internationales (OSI)** développe l'activité dans un souci de rentabilité, de fidélisation des relations adhérents et apporteurs, de recherche d'optimisation des processus de gestion, de suivi des engagements, et de maîtrise des risques pour les produits de la gamme Affacturage en Gestion Déléguée, des opérations spéciales, des produits spécifiques et pour des nouveaux produits.
- **La direction des règlements et services partagés (DRSP)** est un prestataire interne au service de la DRC, d'OSI, ayant pour missions principales le traitement des

règlements, la gestion administrative des contrats et des comptes débiteurs, la gestion des flux débiteurs entrants en support au recouvrement, l'accueil téléphonique des clients et des débiteurs ainsi que des missions de reporting pour les directions clientes. Cette direction permet de fluidifier et de fiabiliser le traitement d'opérations à caractère essentiel tout en optimisant les ressources et en valorisant le travail des collaborateurs qui la composent.

La direction supports regroupe :

- **La direction des systèmes d'information** qui est décrite au chapitre suivant.
- **La direction des ressources humaines** dont la responsabilité est d'être facilitateur dans la gestion des compétences, d'être support et conseil du management, d'être garant du respect de la législation sociale et de vérifier l'ensemble des droits et devoirs issus du contrat de travail.
- **La direction financière** qui tient la comptabilité et gère la trésorerie. Elle assure également la relation avec les commissaires aux comptes, établit les comptes sociaux et consolidés et effectue le reporting réglementaire.
- **La direction programme et qualité** qui regroupe le service Maitrise d'ouvrage applicative qui pilote des projets de différentes natures (organisationnels, stratégiques, financiers) confiés par la Direction Générale ainsi que les services Qualité et Plans et Procédures.

Le secrétariat général assure les relations avec les administrateurs, les actionnaires et les autorités de tutelle. Il organise les réunions du conseil d'administration, du comité d'audit et les assemblées générales. Il comprend également le contrôle interne, qui inclut les fonctions suivantes : contrôle interne permanent, contrôle de la conformité, lutte contre le blanchiment et supervision des missions d'audit périodiques sous-traitées aux auditeurs de la Société Générale.

L'organisation CGA est certifiée ISO 9001 [12] et ISO 14001 [13].

I.2.4 La direction des systèmes d'information

Les missions de la DSI sont nombreuses :

- assurer la continuité des systèmes supportant le métier
- assister les utilisateurs internes et les clients
- prévoir l'évolution des ressources nécessaires
- organiser la direction dans le respect du budget
- participer à l'élaboration du plan de travail, garantir la confidentialité et la sécurité des données
- assurer la maîtrise d'œuvre des évolutions informatiques nécessaires
- gérer le parc machine dans le respect des engagements environnementaux de CGA et du groupe SG.

Pour mener à bien ses missions, la direction des systèmes d'information est structurée selon l'organigramme ci-dessous.

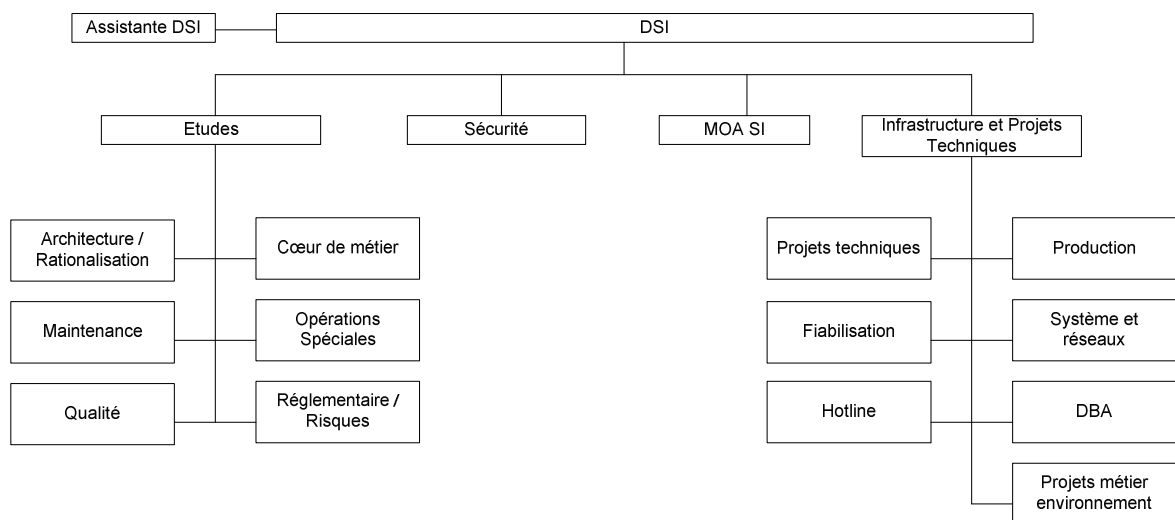


Figure 3 : Organigramme DSI

La DSI est pilotée par une directrice des systèmes d'information et découpée en quatre branches principales :

- **Le service Etudes** (9 collaborateurs internes, 23 collaborateurs externes) : Il est dirigé par un chef de service qui pilote directement les responsables de centre de compétences. Ces centres de compétences sont en charge du développement des projets initiés par la direction générale et formalisés par la direction des programmes. Chaque

centre de compétences est constitué d'un responsable, de chefs de projet et de développeurs et analystes externes. L'adjoint au responsable des études pilote les pôles « Maintenance », « Qualité » et « Architecture / Rationalisation ». Le pôle « Maintenance » prend en charge les demandes d'évolutions et de corrections des applications livrées par les centres de compétences. Le pôle « Qualité » gère les outils de gestion des mini environnements et d'automatisation des tests. Le pôle « Architecture / Rationalisation » intervient dans chaque projet pour orienter les choix d'architecture et piloter l'usage des outils transverses. C'est au sein de ce pôle que je travaille actuellement et que je mène ce projet.

- **Le service Sécurité** (1 collaborateur interne, 2 collaborateurs externes) : Le responsable sécurité met en place les contraintes de sécurité dans le système d'information et vérifie la bonne application de celles-ci. Les aspects sécurité du SI sont étudiés en suivant les axes DICP (Disponibilité – Intégrité – Confidentialité – Preuve).
- **Le service MOA SI** (2 collaborateurs internes, 8 collaborateurs externes) : Il fait l'interface entre la MOA métier et la MOE en traduisant les cahiers des charges en spécifications SI. Les collaborateurs de la MOA SI interviennent également dans la recette technique des projets et applications.
- **Le service Infrastructure et Projets Techniques** (9 collaborateurs internes, 12 collaborateurs externes) : Il est dirigé par un chef de service qui encadre directement les pôles Hotline, Fiabilisation et Projets techniques. La Hotline est une TMA dont la tâche principale est la prise en charge des appels et la résolution des incidents de niveau 1. Le pôle « Fiabilisation » gère la rédaction et l'application des procédures d'exploitation et initie des processus permettant de fiabiliser le SI. Le pôle « Projets techniques » assure le suivi et la réalisation des projets d'infrastructures. Le responsable infrastructure est en charge de l'administrateur des bases de données (DBA) et des pôles « Production », « Système et réseaux » et « Projets métiers environnements », ce dernier mettant à disposition les différents environnements de développement et de recette pour les besoins du service Etudes.

I.2.5 Le système d'information

Dans un contexte de marché très porteur, le système d'information de CGA s'est considérablement développé ces dernières années. Eu égard à l'élargissement de son offre commerciale et au nombre croissant de produits proposés à sa clientèle, le système d'information de CGA se caractérise par sa lourdeur et sa complexité.

- **L'infrastructure**

Le système d'information de CGA repose principalement sur une plateforme IBM Pureflex physiquement localisée chez un hébergeur. Ce serveur supporte les partitions AS400 (production, homologation et développement) qui contiennent les applications représentant le cœur du SI ainsi que les serveurs Intel virtualisés les plus sensibles (serveurs d'application, serveurs de fichiers, serveurs d'impression, serveurs de messagerie...). Certains serveurs Intel moins sensibles sont toujours localisés au siège social de CGA mais une opération de virtualisation et de migration des serveurs physiques est actuellement en cours. L'exploitation de l'infrastructure est pilotée par de nombreux outils : ALDON LMI et LME (outil de livraison et de versioning permettant d'industrialiser les mises en production), Robot400 (scheduler central permettant de déclencher les traitements batchs de manière événementielle ou planifiée), Bridge400 (aiguilleur de flux permettant le transfert des fichiers entre les différents serveurs du SI), Gateway (passerelle permettant de gérer les échanges de fichiers entre CGA et l'extérieur), Tango04 (outil de supervision permettant de monitorer l'exécution des traitements).

En termes de postes client, un chantier de mise à jour est également en cours pour migrer les postes de Windows XP vers Windows 7 en respectant les contraintes édictées par le groupe Société Générale.

- **Le patrimoine applicatif**

Le cœur de métier est composé du progiciel d'affacturage Aquarius auquel est rattachée la quasi-totalité des applications « connexes ». Celles-ci utilisent les données d'Aquarius soit en lecture, soit en écriture. 22 applications sont jugées critiques car elles contribuent aux services-clés distribués par CGA : le financement, la délivrance des lignes crédit, la relance

et le recouvrement des créances, l'encaissement des règlements débiteurs, le portail internet client.

Afin d'accroître l'efficacité opérationnelle des exploitants, notamment gestionnaires et chargés de clientèle, CGA a automatisé le plus possible, au fil des années, toutes ces activités. Ainsi, le SI de CGA est constitué à ce jour d'une quinzaine de progiciels du marché et d'une centaine d'applications internes développées via une quinzaine de technologies différentes (Java, RPG, Cobol, Lansa, Windev, Webdev, ..).

- **Les flux de données**

L'activité affacturage se caractérise par l'échange de flux commerciaux et financiers. De nombreuses applications lisent, transforment ou créent des flux, qui font évoluer le statut des données inhérentes aux contrats, tiers, et items. Le traitement de flux de données est donc au cœur du métier de CGA.

Un grand nombre de fichiers sont échangés quotidiennement. Ces échanges se font aussi bien avec l'extérieur (clients, fournisseurs et partenaires de CGA) qu'à l'intérieur du SI (entre les différentes briques applicatives). Les fichiers provenant (ou à destination) de l'extérieur sont transférés via divers canaux : un site extranet transactionnel, une passerelle et un gestionnaire CFT. Quelque soit leur provenance (interne ou externe) tous les fichiers transitent par le planificateur de flux installé sur la plateforme centrale iSeries. Ces transferts de fichiers sont soit événementiels, soit déclenchés par le planificateur de tâches installé sur la plateforme centrale iSeries.

En termes de volumétrie, 4000 fichiers transitent quotidiennement par le planificateur de flux (avec des pics à 7000 fichiers en fin de mois). Certains fichiers échangés entre les applications représentent des volumes conséquents (ex : le chargement quotidien des items entre le logiciel cœur de métier et le site transactionnel représente 700 000 enregistrements en moyenne quotidiennement). Les types de fichiers échangés sont multiples : fichiers CSV avec séparateurs, fichiers texte au format fixe, fichiers XML, fichiers PDF, fichiers image (TIFF...) ... Les formats des fichiers échangés sont majoritairement définis par CGA (fichiers de factures, fichiers retours...). Des formats spécifiques aux clients ainsi que des formats normalisés (CFONB, SEPA...) sont également échangés.

I.2.6 Le schéma directeur

Le système d'information de CGA s'est construit progressivement depuis plus de dix ans dans une logique « business » grâce à la compétence des acteurs de CGA. A ce jour, le niveau de complexité a atteint ses limites, et la DSI se doit d'industrialiser et moderniser le système d'information, afin d'aider CGA à poursuivre sa croissance. Dans cette optique, CGA a entamé une démarche de schéma directeur.

Cette démarche initiée en 2013 avec un cabinet spécialisé dans la rationalisation des systèmes d'information a permis dans un premier temps de constater les limites dans le SI actuel de CGA vis-à-vis de ses ambitions de développements :

- l'automatisation des modes de fonctionnement est au maximum de ce que permettent les systèmes actuels. Cependant des tâches critiques ne sont pas automatisées. Elles exposent l'entreprise à des risques financiers et génèrent des impacts opérationnels multiples pour l'entreprise.
- le système actuel comporte de nombreuses applications dupliquées et fortement couplées qui limitent l'évolutivité fonctionnelle du SI. Il en résulte un accroissement des coûts de maintenance et d'évolution du fait des nombreux impacts à gérer.
- le foisonnement des technologies, pour assurer le fonctionnement du patrimoine applicatif complexifie le SI. Il implique également le maintien de compétences informatiques multiples pour faire face à l'hétérogénéité applicative et technique des outils.

Cette analyse a permis, dans un second temps, de déterminer les objectifs suivants qui permettront à CGA de reconfigurer son SI et de s'engager sur une voie de l'industrialisation :

- **favoriser l'industrialisation et la dématérialisation de solutions standardisées.**
Dans cette optique, deux architectes ont rejoint la DSI de CGA pour intervenir dans les phases d'avant-projet et ainsi uniformiser les décisions prises en termes d'architecture logique et technique et notamment privilégier le choix de solutions standardisées plutôt que le recours à des développements spécifiques.

- **moderniser le SI et intégrer les nouveaux usages et les besoins d'agilité.** Pour répondre à ce besoin, de nombreux projets sont déjà en cours afin de remplacer les éléments vieillissants du SI (planificateur, aiguilleur de flux, superviseur...).
- **améliorer l'efficacité opérationnelle.** La réorganisation du service Etudes va notamment être entreprise dans ce sens avec la création d'un pôle maintenance qui prendra en charge la maintenance évolutive et applicative et permettra aux autres pôles de se concentrer uniquement sur les nouveaux projets.
- **rationaliser le SI et en réduire les coûts.** Cet objectif va notamment être réalisé grâce à la disparition de l'ancien progiciel et le dé-commissionnement des applications connexes liées à celui-ci. Une réflexion est également menée pour maximiser l'utilisation des fonctionnalités existantes dans le progiciel cœur de métier et ainsi éradiquer les applications connexes inutiles.
- **réviser les principes de gouvernance et de gestion de projets** entre les équipes DSI (MOE), Direction de Programme (MOA) et les lignes métier de CGA. Dans cette optique, des interlocuteurs clés ont été identifiés au sein des différentes directions de CGA. Un rapprochement entre MOE et MOA est également en cours avec la réorganisation de la DPR et la création d'un service MOA-DSI.

II La problématique

II.1 Les enjeux

Dans le cadre du schéma directeur et notamment de son objectif de rationalisation du SI, un portefeuille de projets a été identifié. Le projet de mise en place d'une solution d'intégration des données fait partie de ce portefeuille. Il se justifie particulièrement au vu de l'activité principale de CGA qui se caractérise par la gestion de flux. Suite à une étude sur les types de solutions existantes (cf. chapitre suivant), CGA souhaite donc acquérir un ETL en mode « moteur », c'est-à-dire un logiciel disposant d'un moteur de transformation des données capable d'assurer des transformations complexes de données indépendamment des langages de manipulation de données des SGBD sources et cibles. Cet ETL sera apte à paralléliser massivement des traitements sur des données issues de sites différents et hétérogènes.

L'ETL devrait nous permettre la synchronisation massive d'informations d'une banque de données vers une autre. L'ETL permettra d'éviter la réalisation de programmes batches répétitifs, souvent semblables, dont il faut également assurer la maintenance. Le principe étant que l'intégration d'un nouveau flux de données ne requiert aucun développement et s'opère par une simple configuration interactive. On choisit les éléments de données dans le référentiel source, on indique les transformations qu'ils doivent subir, et on précise la destination de la donnée dans la cible.

Dans le contexte de CGA, l'ETL doit permettre notamment de s'affranchir des développements inhérents aux mises en place de flux avec les nouveaux clients. Il doit également permettre de faire disparaître certaines briques logicielles (ETL existant, outils de chargement inter-applications, convertisseur de formats ...) et de simplifier des chaînes de traitement existantes.

La solution retenue deviendra une brique centrale du SI de CGA et devra, de ce fait, être prise en compte dans chaque nouveau projet lors de la définition de l'architecture logicielle. Enfin, ce nouvel outil nous permettra également de moderniser notre SI et de pouvoir intégrer de nouveaux profils dans nos équipes en s'affranchissant des compétences AS400 qui se raréfient.

Le projet qui m'est confié sera donc désigné chez CGA sous le nom « Mise en place d'un ETL ».

Ces principales phases sont :

- choisir la solution ETL la plus adaptée au contexte de CGA.
- installer la solution retenue dans les environnements de production et de non-production.
- intégrer cette solution en mettant en place les interactions nécessaires avec les autres briques techniques du SI.
- développer les premiers flux inclus dans les projets contraints.
- définir les bonnes pratiques liées à l'utilisation de l'ETL en fournissant notamment des templates et des méthodes à suivre aux développeurs.

Aucune contrainte de coûts n'est définie pour ce projet. En revanche, une contrainte de délai est imposée puisque le premier flux identifié pour être traité par l'ETL concerne un nouveau client majeur dont le contrat démarre en juin 2014.

En termes de méthodologie, je suivrai la méthode de conduite de projet existante chez CGA (cf. chapitre VI) que j'enrichirai grâce aux connaissances acquises lors de mon cursus au CNAM avec notamment les préconisations du PMBOK [01].

II.2 Focus sur les solutions d'intégration de données

J'ai mené dans une phase d'avant-projet un rapide état des lieux sur les solutions d'intégration de données existantes sur le marché. J'ai également profité de cette phase pour installer un ETL open-source (Talend) sur mon poste afin d'appréhender les possibilités offertes par un ETL.

II.2.1 La définition des ETL

« Extract – Transform – Load » est connu sous le terme ETL. Il s'agit d'une technologie permettant de construire de façon simple des processus qui permettent d'extraire des données de systèmes hétérogènes, de les transformer, puis de les charger dans un système cible.

Les solutions ETL reposent sur :

- des connecteurs servant à exporter ou importer les données
- des transformateurs (agrégateurs, filtres, convertisseurs...) qui manipulent les données
- des mises en correspondance (des mappages)

A l'origine, les ETL sont apparues pour le chargement régulier de données agrégées dans les entrepôts de données de l'informatique décisionnelle (Datawarehouse), avant de se diversifier vers la problématique plus globale d'intégration de données.

II.2.2 Les normes et standards

Il n'y a pas de standards propres à l'ETL, les standards utilisés sont ceux liés :

- à l'accès aux bases de données : essentiellement ODBC/JDBC même si les accès natifs, plus performants, sont privilégiés
- aux bus de communication : DCOM (Microsoft), EJB (plateforme JEE) ou MOM (Messagerie Orientées Messages)
- aux normes du transfert de fichiers : FTP, sFTP...

- aux formats des échanges : XML, SWIFT...

II.2.3 Les tendances

Les technologies ETL ont évolué. L'objectif des technologies ETL s'est désormais élargi à l'intégration de l'entreprise par ses données.

Les ETL de contenu permettent de manipuler des données non structurées tels que les dossiers et les documents. Ces solutions sont utilisées pour des projets de migration de documents (par exemple, lors d'une migration d'une application GED). Leur champ d'application peut également s'étendre à des projets d'archivage électronique.

Les ETL s'orientent davantage vers des problématiques globales d'intégration des données, temps réel et récupération des données directement en production. Ce sont des problématiques de type Entreprise Application Integration (EAI) ou Entreprise Information Integration (EII).

- **L'orientation EAI**

L'ETL s'oriente de plus en plus vers des alimentations de type « temps réel », ce qui le rapproche à ce titre des EAI. Et si les discours marketing des éditeurs portent souvent à confusion sur des offres qui feraient tout, il demeure toutefois des différences fondamentales :

- la vocation de l'ETL est de migrer de gros volume de données d'une base à l'autre le plus souvent en mode batch et à destination de systèmes essentiellement décisionnels.
- la vocation de l'EAI est de synchroniser, via un « bus applicatif », des applicatifs métiers entre eux via des échanges synchrones de faibles volumes de données. L'EAI rentre dans la couche transactionnelle. Les EAI s'appuient sur un référentiel de règles métier alors que les ETL s'attachent à faire correspondre des schémas de données non compatibles.

- **L'orientation EII**

L'EII est un outil, doté de fonction de mapping, visant l'invocation en mode synchrone de requêtes distribuées entre plusieurs applications métiers hétérogènes. Il s'agit ainsi d'être en mesure de récupérer des données de production « à la volée » puis de les comparer à des données historiques qui sont elles hébergées dans le datawarehouse.

Les éditeurs d'EII sont pour l'instant distincts de ceux de l'ETL puisqu'ils prônent l'accès direct aux données de production contrairement à ceux d'ETL dont la raison d'être est de construire une base de datawarehouse intermédiaire. L'association d'outils EII au décisionnel laisse cependant entendre qu'à terme, les outils d'analyse interrogeront directement les bases de production ce qui « court-circuitera » les étapes de consolidation dans un entrepôt de données de type datawarehouse. Pour l'heure, cette perspective semble encore très lointaine et ce pour deux raisons :

- les exigences de cohérence des données : l'exploitation de sources hétérogènes implique que les définitions soient partagées et homogènes
- les performances : il est délicat de faire cohabiter à la fois une activité transactionnelle (qui bloque des enregistrements de base de données) et une activité décisionnelle qui parcourt de larges parties de la base de données.

Les éditeurs ETL qui s'étaient engouffrés sur le marché des applications analytiques ont depuis fortement désinvesti dans le domaine. Ils ont préféré laisser ce créneau aux acteurs de la restitution qui eux ont évolué vers des plateformes décisionnelles intégrées et proches de la connaissance métier. Les éditeurs ETL ont au mieux conservé des environnements pré configurés mais ont cessé tout développement d'applications clés en main.

Les éditeurs investissent le volet Qualité de données. A l'origine destinés à des besoins spécifiques tels que les campagnes marketing, cette fonctionnalité est de plus en plus souvent sollicitée pour améliorer les informations échangées entre les progiciels.

II.2.4 Les outils et acteurs

Le marché de l'ETL est porté par les efforts des entreprises pour gagner à la fois en qualité et sécurité (gouvernance des données), en rapidité de mise en œuvre (Time To Market) et en agilité (optimisation continue des processus métier).

Le marché de l'ETL se concentre et se restructure autour des nouvelles fonctionnalités (cf. chap. II.2.3). De fait, les quelques éditeurs purs « players » sont de plus en plus concurrencés par les éditeurs décisionnels, base de données, voir middleware qui intègrent l'ETL dans leurs suites intégrées.

Les outils purement dédiés ETL vont probablement devenir sur les prochaines années des outils de niche, restreints à l'ultra performance et à l'interopérabilité ou aux besoins les plus basiques.

Seuls les plus gros éditeurs (IBM, Informatica, SAP, Oracle et Microsoft) sont en mesure de proposer des solutions d'intégration de données et non plus seulement d'ETL.

Des solutions Open Source existent également (Talend...) mais une richesse fonctionnelle en retrait et un support de moindre qualité restreignent leur utilisation. Ceux-ci sont aujourd'hui utilisés dans des niches, des contextes d'exigences fonctionnelles et de performances limitées, de périmètre limité à un projet (et non l'entreprise entière) et enfin et surtout où le coût est le facteur primordiale.



Figure 4 : Magic Quadrant – Outil d'intégration de données

Le Magic Quadrant de Gartner nous fournit une vision des acteurs du marché des outils d'intégration de données.

III L'acquisition d'une solution ETL

La phase d'acquisition est la phase initiale du projet. Elle consiste à acquérir une solution d'intégration des flux de données répondant aux attentes de CGA.

Cette phase a été menée en s'inspirant des bonnes pratiques décrites dans le domaine du management des approvisionnements du PMBOK adaptées aux pratiques existantes dans l'organisation CGA et au contexte du projet.

III.1 L'appel d'offre

Selon le PMBOK, le processus « Planifier les approvisionnements » est « le processus qui consiste à documenter les décisions d'approvisionnements du projet, à spécifier les approches et à identifier les vendeurs potentiels ».

Les données d'entrée de ce processus sont les exigences exprimées dans le schéma directeur du système d'information CGA notamment dans la définition de l'enjeu principal d'optimisation des coûts dont l'objectif est de rationaliser les coûts informatiques (projets et/ou maintenance) à travers la mise en place de normes et de processus industrialisés. Les accords de partenariat entre CGA ou le groupe Société Générale et les fournisseurs potentiels seront également influents dans ce processus.

Les techniques utilisées pour produire les livrables sont notamment le jugement d'experts (à travers les interviews du responsable sécurité et des collaborateurs ayant expérimenté les solutions d'intégration de données dans leurs expériences précédentes) et l'analyse de l'état de l'art des ETL.

Ce processus a abouti à la rédaction d'un document d'approvisionnement (un appel d'offre) transmis aux différents consultés et à la détermination des critères de sélection des réponses.

L'appel d'offre décrit une vision macroscopique des attentes de CGA vis-à-vis de la solution apportée par le fournisseur. Cet appel d'offre est constitué de :

- **une description du contexte CGA** (cf. Chapitre I)
- **une description du système d'information existant** (cf. Chapitre I.2.5)
- **un focus sur le traitement des flux de données**

Ce focus permet notamment d'appréhender l'intérêt attendu de la solution ETL au regard de la complexité existante dans le SI de CGA.

Le schéma ci-dessous représente les différents types de flux existants actuellement.

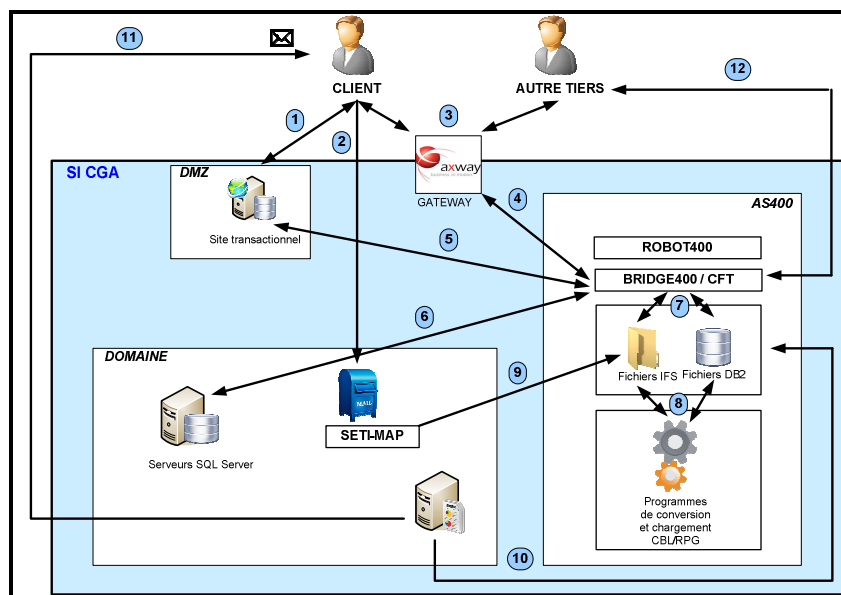


Figure 5 : Schéma simplifié des flux du SI CGA actuel

- ① Les clients peuvent transmettre leurs fichiers (bordereaux, images...) à CGA par l'intermédiaire d'un upload (ou d'une saisie) sur le site transactionnel. Ils peuvent également recevoir des fichiers, que CGA leur met à disposition, en les téléchargeant sur le site transactionnel (états PDF, fichiers retour CSV).
- ② Les clients peuvent transmettre leurs fichiers (factures, avoirs, débiteurs...) à CGA par l'envoi d'un mail à une boîte aux lettres générique.
- ③ Les clients et autres tiers (prestataires, banques...) peuvent transmettre et recevoir des fichiers par liaison sécurisée au travers de la Gateway de CGA.
- ④ La Gateway communique avec l'AS400 via CFT (encapsulé dans le planificateur de flux).
- ⑤ Le site transactionnel communique avec l'AS400 (transfert de fichier et intégration dans la base SQL Server) via le planificateur de flux.
- ⑥ Les fichiers de chargement de base de données SQL Server (datawarehouse, applications connexes...) sont construits à partir des fichiers de l'application métier (DB2

AS400) par des programmes spécifiques sur l'AS400 puis transférer via le planificateur de flux.

⑦ Les fichiers AS400 (IFS et DB2) sont déplacés à l'intérieur de l'AS400 via le planificateur de flux.

⑧ Les fichiers AS400 (IFS et DB2) sont manipulés (transformation, conversion) par des programmes spécifiques développés par la DSI CGA.

⑨ L'ETL Seti-Map récupère les messages dans la boîte aux lettres générique, extrait les pièces jointes contenues, les décompresse, leur applique une transformation et les dépose sur l'IFS.

⑩ Les fichiers DB2 AS400 sont utilisés par des programmes d'édition de rapports de traitement.

⑪ Les rapports générés à l'étape 10 sont ensuite transmis par mail aux clients.

⑫ Certains tiers (prestataires...) peuvent transmettre et recevoir des fichiers directement via CFT (encapsulé dans le planificateur de flux).

Le schéma ci-dessous, représentant la cible à atteindre, décrit l'objectif de centralisation et de simplification des différents flux grâce à l'ETL. Ce dernier s'intégrera en tant que brique centrale du SI permettant la gestion des transformations de flux entre les applications et les SGBD. L'ETL pourra également intervenir dans des opérations de nettoyage des données afin d'augmenter la qualité de celles-ci et dans des opérations ponctuelles (migration, chargement bulk...)

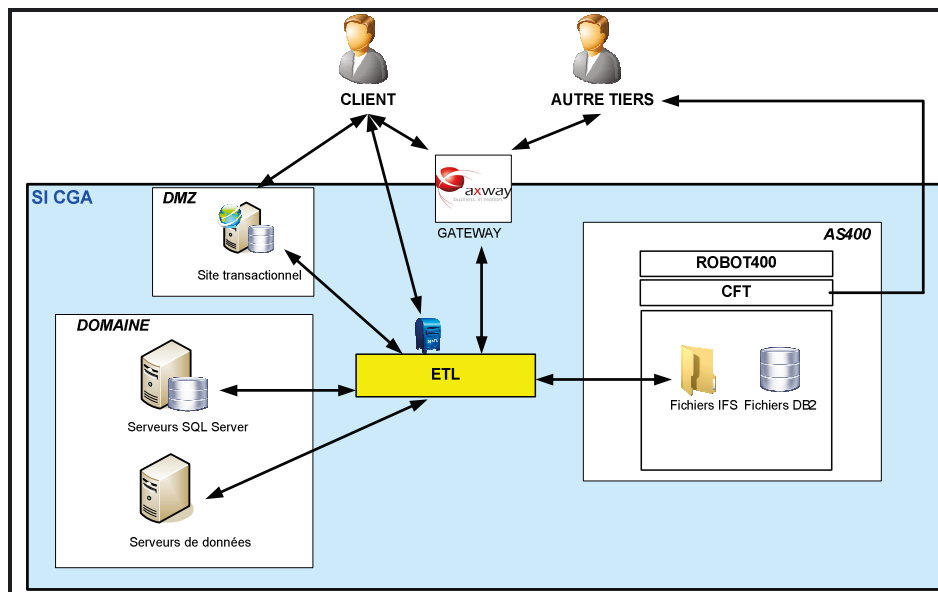


Figure 6 : Schéma simplifié des flux du SI CGA intégrant l'ETL

- **une description des objectifs de l'acquisition d'un ETL** (cf. Chapitre II)
- **les fonctionnalités attendues pour l'outil ETL**

L'outil ETL proposé sera jugé notamment sur les fonctionnalités suivantes :

Plateformes/SGBD supportés

Les plateformes du système d'information CGA étant iSeries et Windows, la solution devra pouvoir être installée sur l'une ou l'autre de ces plateformes. Dans le cas d'une installation sur iSeries, la base de données D2 devra être supportée. Dans le cas d'une installation sur Windows, la base de données SQL Server devra être supportée.

Fichiers supportés

Les échanges et conversions de fichiers plats sont nombreux au sein du SI CGA et concernent principalement des fichiers TXT, CSV, XML, XLS, PDF, TIFF, JPEG... Ces types de fichiers devront donc être supportés. Le format des fichiers peut être mono ou multi-format.

Accès aux données

Les SGBD suivants sont utilisés dans le SI de CGA et devront donc être supportés par la solution : DB2, SQL Server, MySQL, Oracle. Dans un souci d'évolutivité, le support de SGBD supplémentaires sera apprécié. La solution devra également proposer des connecteurs pour accéder aux données contenues dans des applications hétérogènes

(LDAP, Sage...). Une solution compatible avec les sources de données ODBC serait un atout non négligeable.

Interface graphique intuitive

L'outil n'étant pas uniquement manipulé par des développeurs, sa prise en main et sa manipulation devront être aisées. Les traitements devront pouvoir être exécutés par cette interface graphique mais également en ligne de commande ou via des API incluses dans la solution.

Processus

L'outil doit permettre de séquencer plusieurs transformations. Ces processus devront pouvoir être déclenchés par message (JMS, MQSeries, Mail...) ou par polling (polling directory, polling base de données, déclenchement par trigger...).

Exploitation et Administration

Les tâches d'administration standard devront être disponibles : reprise sur incident, gestion des erreurs, visualisation de l'historique des traitements, pistes d'audit ...

La gestion des droits d'accès à la solution devra être explicitée par le consulté (synchronisation avec un annuaire LDAP ou gestion autonome). Dans le cas d'une gestion autonome, les notions de type d'utilisateurs, complexité des mots de passe, péremption, historique des connexions, blocage des comptes devront être gérées.

Le produit devra intégrer des dispositifs d'alerte pour pouvoir idéalement être intégré dans une solution de supervision globale.

Sécurité

Les aspects « sécurité » de la solution devront faire l'objet d'une description détaillée de la part du consulté. Les notions de disponibilité, intégrité, confidentialité et preuve devront être abordées. Le consulté devra faire part de ces préconisations en termes d'architecture (haute disponibilité, PRA, sauvegarde / restauration). La gestion des licences doit également être abordée.

Caractéristiques spécifiques

La solution devra fournir des traitements spécifiques tels que : la manipulation de courriels (récupération et envoi de mails), la décompression ainsi que le cryptage/décryptage de fichiers archives (zip, tar...).

La gestion d'environnement devra également être gérée (Développement, Homologation, Production). Les configurations créées devront être transportables entre ces environnements. Les transferts de fichier devront utiliser des protocoles connus par CGA : FTP, sFTP, CFT... La solution devra s'interfacer dans l'ordonnanceur (scheduler) de jobs existant. Les transformations disponibles devront être nombreuses (jointures, transcodages ...) et le consulté devra préciser dans quel langage celles-ci sont codées (SQL, langage propriétaire, langage script...).

- **les services attendus de la part du consulté**

Le consulté devra être en mesure de fournir à CGA une assistance à la mise en place de la solution dans le SI de CGA en termes d'installation et de paramétrage des pré-requis.

Le consulté devra également être en mesure de fournir à CGA une assistance à l'utilisation en termes de conseil et d'aide au paramétrage. Le support technique devra également être disponible et réactif.

Enfin, le consulté devra proposer une formation à l'outil proposé.

- **les instructions générales concernant la réponse attendue**

Les instructions concernant les engagements de CGA et du consulté, la confidentialité et le contenu des réponses sont décrites.

Ce processus a également permis de déterminer les critères de sélection des sources. Ainsi CGA notera et classera les consultés sur la base de critères qualitatifs et attribuera la plus grande priorité aux critères suivants : adéquation de la solution aux besoins exprimés, simplicité de la mise en œuvre, coût global (solution et services) et qualité de la réponse.

La mise en place d'un outil ETL nous permettra de simplifier un certain nombre de scénarios existants actuellement dans le SI de CGA. Les scénarios suivants sont des traitements courants dans les développements actuels et futurs du SI de CGA, ils sont décrits dans un document joint à l'appel d'offre pour éclairer les consultés sur nos besoins détaillés :

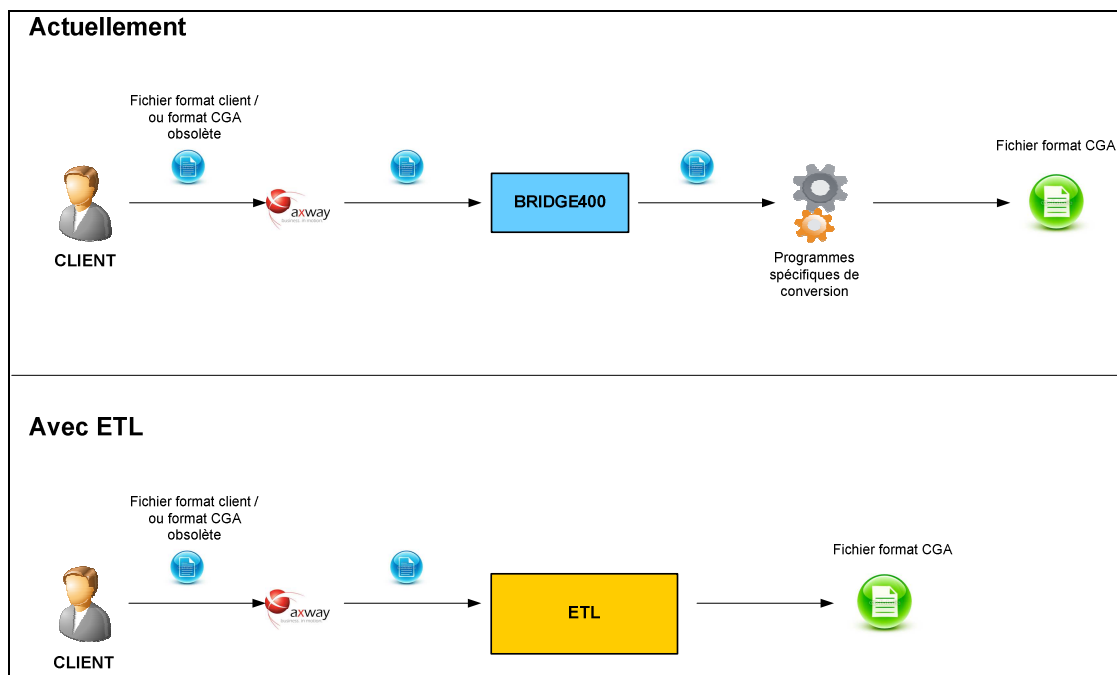


Figure 7 : Cas d'utilisation : Réception de fichiers clients non standards

CGA échange de nombreux fichiers avec ses clients (factures, avoirs, débiteurs...). Les formats de ces fichiers sont déterminés par CGA mais certains clients importants refusent de s'y soumettre et préfèrent nous transmettre un format personnel. Les changements de version des formats de fichier posent également problème car nous ne pouvons demander aux clients « historiques » de changer leur format. Donc, pour un même type de fichier, nous recevons différents formats (versions différentes). Pour gérer l'ensemble de ces problématiques, un grand nombre de programmes spécifiques ont été développés, la maintenance et le développement de ces programmes sont lourds.

L'ETL a pour but de simplifier et d'optimiser cette étape en permettant de paramétrer ces conversions en s'affranchissant de toute programmation. Il apportera également une plus grande souplesse vis-à-vis des formats échangés.

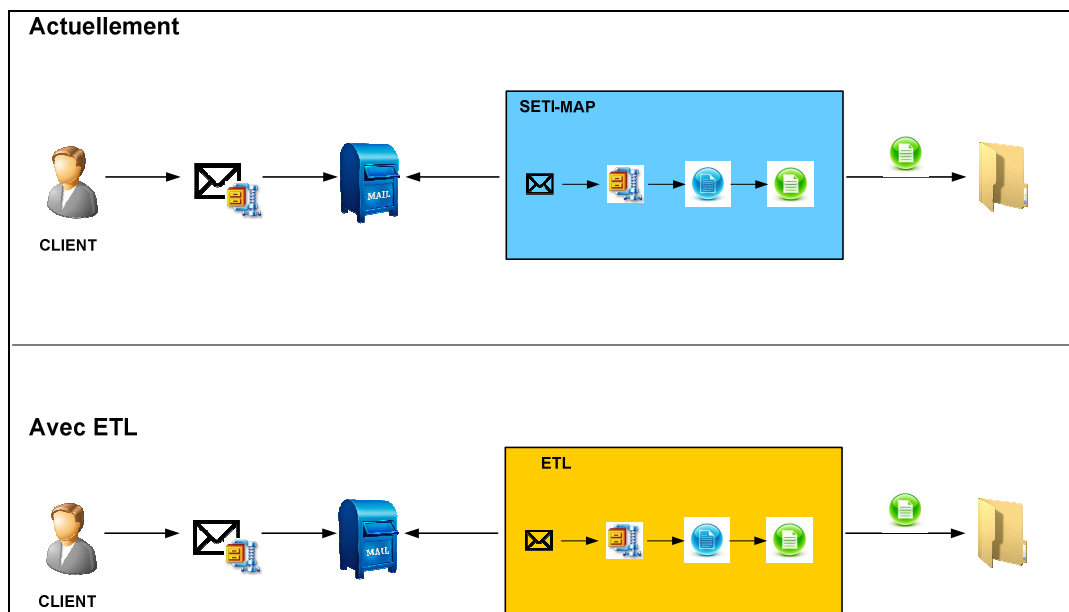


Figure 8 : Cas d'utilisation : Réception de fichiers clients par mail

Certains clients nous transmettent leurs fichiers en tant que pièces jointes zipées dans un mail. CGA dispose d'un outil (SETI-Map) permettant de scruter une boîte aux lettres générique, d'extraire de chaque mail reçu les pièces jointes, de décompresser ces pièces jointes puis de convertir les fichiers obtenus en y appliquant des règles de mappage. L'ETL a pour but de remplacer la solution SETI-Map en permettant donc les mêmes fonctionnalités (scrutation BAL, décompression, conversion).

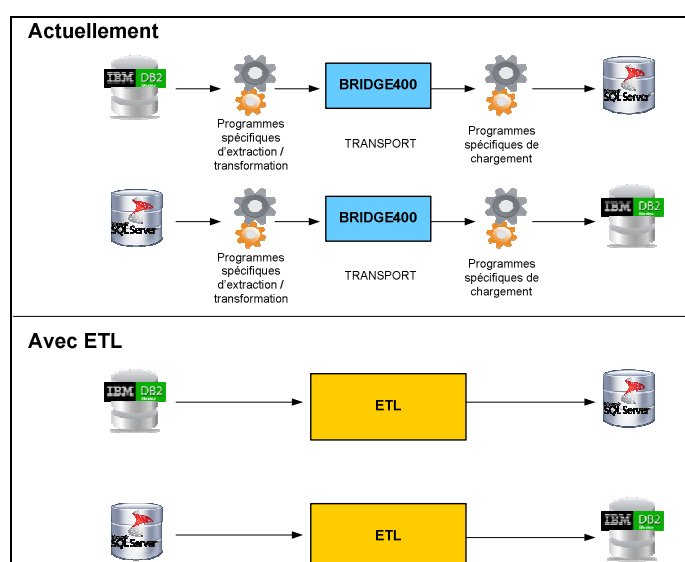


Figure 9 : Cas d'utilisation : Chargement inter-applications

La plateforme principale de CGA est un iSeries OS400 V6R1. L'application métier et de nombreuses applications connexes sont hébergées sur cette plateforme sous forme de

fichiers DB2 AS400. D'autres applications connexes sont hébergées sur des plateformes Windows avec des bases de données SQL Server. De nombreux échanges ont lieu entre ces différentes applications. L'ETL a vocation à fluidifier les échanges entre les applications DB2 AS400 et les applications SQL Server et inversement.

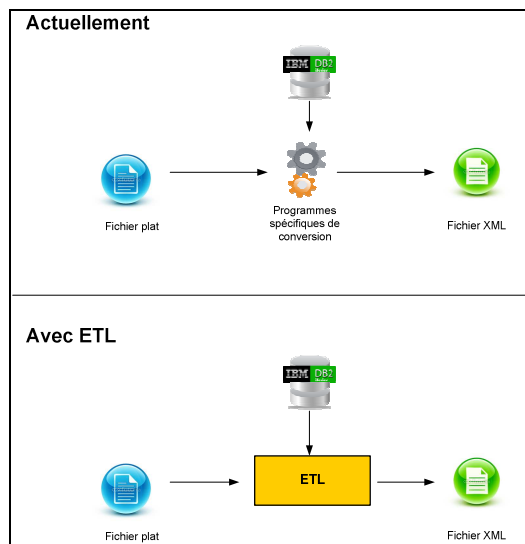


Figure 10 : Cas d'utilisation : Conversions complexes avec accès à un référentiel

Des conversions complexes entre un format source plat et un format cible XML (par exemple) sont réalisées par des programmes spécifiques et complexes de conversion. Ces programmes enrichissent le fichier cible avec des données contenues dans des référentiels DB2 AS400. L'ETL a vocation à remplacer ces développements en permettant notamment l'enrichissement par l'accès aux référentiels CGA.

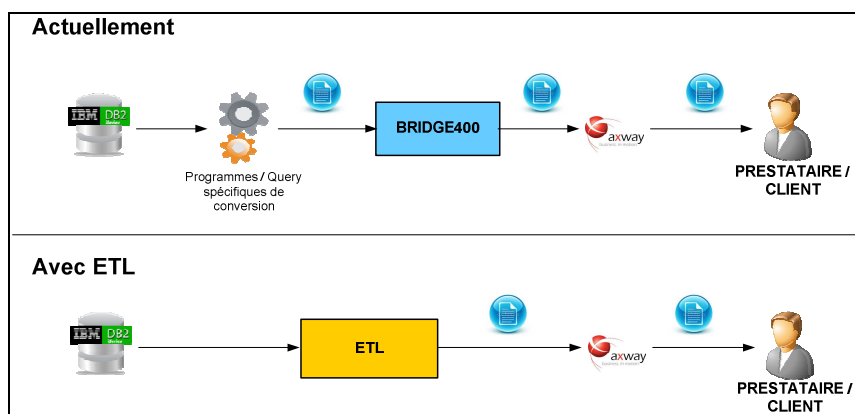


Figure 11 : Cas d'utilisation : Extraction des données métiers vers un fichier plat

De nombreuses applications sont réalisées à partir des données de la plateforme principale iSeries sous forme de programmes spécifiques ou queries. Ces extractions sont produites au format CSV pour être transmises aux prestataires et clients mais également en interne aux

utilisateurs du SI. L'ETL a vocation à remplacer un maximum de ces développements/queries.

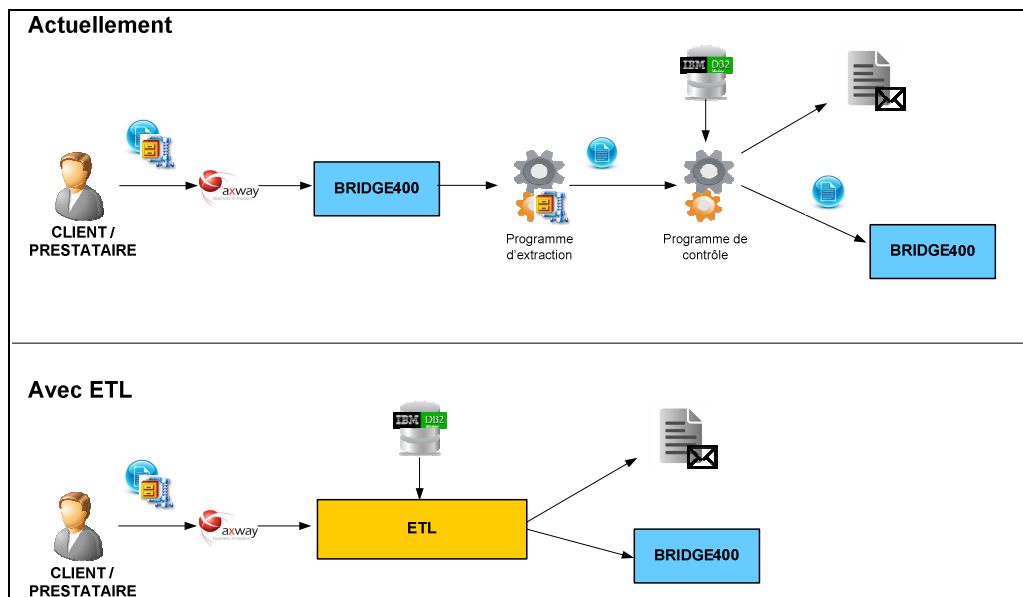


Figure 12 : Cas d'utilisation : Contrôles et traitements spécifiques de fichiers reçus

Certains fichiers nécessitent des traitements spécifiques : décompression, contrôle de contenu du zip, émission d'un rapport listant les erreurs de format dans le fichier reçu... L'ETL devra pouvoir réaliser ces traitements spécifiques et ainsi remplacer les programmes spécifiques développés pour ces besoins.

Dans le cadre d'opérations ponctuelles de migration ou de chargement massif, l'ETL sera amené à intervenir. Il permettra de simplifier les étapes suivantes : extraction de données, nettoyage des données, analyse des données, transformation des données, chargement des données dans l'application cible.

Ces opérations sont exceptionnelles (mise en place de la nouvelle version d'une application, remplacement d'une application existante par une autre, acquisition d'une nouvelle application) mais souvent lourdes en termes de charge et non réutilisables (one-shot).

III.2 La réponse au besoin et le choix de la solution

Quatre fournisseurs sont consultés dans le cadre du projet. Ce panel a été élaboré en fonction des préconisations du groupe SG, de la réputation de ces fournisseurs dans le domaine étudié et des relations que nous entretenons déjà avec ceux-ci.

Selon le PMBOK, le processus « Procéder aux approvisionnements » est « le processus qui consiste à obtenir les réponses des vendeurs, à sélectionner un vendeur et à attribuer un contrat ».

Les données d'entrée principale de ce processus ont été les réponses des consultés à notre appel d'offre. Ce processus particulièrement décisif dans la réussite du projet a été réalisé en plusieurs étapes :

- **Présentation des solutions** : Avant de rendre leur réponse à notre appel d'offre, les fournisseurs sont venus nous présenter les fonctionnalités standard de leurs solutions.
- **Analyse des réponses** : La réception des réponses à l'appel d'offre a été le déclencheur d'une première phase d'analyse consistant en la lecture attentive des réponses et quelques entretiens téléphoniques avec les consultés pour apporter des éclaircissements. La conclusion de cette étape a été la sélection d'une short-list de laquelle l'un des fournisseurs a disparu car sa solution ne couvrait que trop partiellement les attentes exprimées.
- **Soutenances** : Chaque fournisseur est venu chez CGA nous présenter sa réponse à notre appel d'offre en essayant de nous démontrer concrètement la résolution des problématiques exprimées dans l'expression de besoins.
- **Analyse finale des réponses** : Cette analyse est l'étape primordiale qui nous amènera à préconiser une solution vis-à-vis de ces concurrentes. Pour réaliser cette analyse, des techniques d'analyse comparative ont été utilisées. Notre analyse a été menée sur les quatre critères identifiés lors du processus de planification des approvisionnements.

L'adéquation de la solution aux besoins est évaluée par l'utilisation d'une grille récapitulant les besoins identifiés et la réponse des fournisseurs à chaque besoin (cf. Annexe 1).

Les coûts des différentes solutions proposées sont compilés dans un document d'analyse des coûts (cf. Annexe 2).

La simplicité de mise en œuvre est évaluée au travers d'entretiens complémentaires avec les fournisseurs ainsi que l'évaluation des durées et méthodes d'implémentation proposées.

La qualité de la réponse est une évaluation plus subjective qui regroupe notamment le respect des instructions données lors de l'appel d'offre, la prestation lors de la soutenance et les relations avec les interlocuteurs du fournisseur.

Le résultat de ces travaux aboutit à un classement valorisé des réponses (cf. Annexe 3) et un diagramme expliquant notre préconisation.

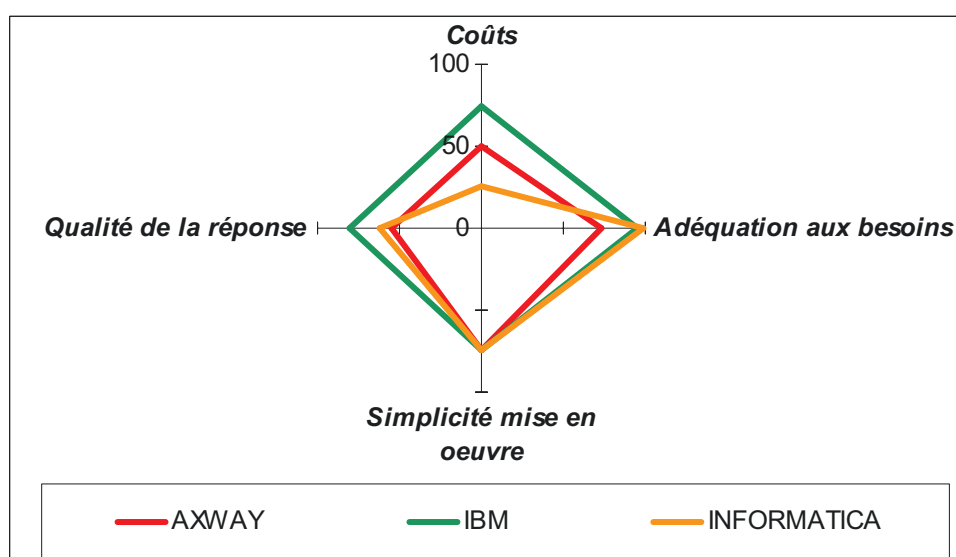


Figure 13 : Diagramme d'évaluation des fournisseurs

- **Présentation à la Direction Générale** : Le résultat des analyses est argumenté devant la direction générale qui accepte notre préconisation concernant le choix de la solution la plus adaptée à nos besoins et notre contexte. Nous lui fournissons également un business case quinquennal (cf. Annexe 4) pour évaluer le ROI prévisionnel inhérent à la mise en place de la solution au cours des prochaines années. L'élaboration de ce business case est basé sur des hypothèses provenant de l'expérience des projets chez CGA et de l'expérience du fournisseur sur sa solution. Ce business case met en avant les gains espérés en termes de nouveau développement (plus rapide), de maintenance applicative (plus simple), de

maintenance logicielle (économie sur les briques applicatives supprimées) et d'automatisation (intégration des travaux de tests de fichiers réalisés manuellement par un ETP).

- **Décision finale** : La décision finale a été entérinée par un document d'engagement de dépenses signé par la direction générale.

La donnée de sortie majeure de ce processus est le contrat d'approvisionnement signé avec le fournisseur retenu.

III.3 La présentation générale de la solution retenue

La solution retenue est Infosphere Datastage d'IBM. Cette solution est un élément de la suite Information Server (plateforme d'intégration des données) qui est constituée de 4 modules reposant sur un socle technique unifié :

- Information Analyzer : compréhension automatisée de données pour lever le voile sur le contenu des données sources et leur structure.
- QualityStage : nettoyage et amélioration de la qualité des données pour identifier, corriger, faire correspondre, standardiser et réconcilier des données imprécises ou redondantes.
- Federation Server : fédération ou virtualisation d'accès aux données afin de rendre accessible chaque information / donnée à tous (utilisateurs, applications, SGBD...)
- DataStage : transformation et distribution des données pour obtenir les données quelles que soient leur complexité et leur source, formatées selon les exigences et distribuées à n'importe quelle cible, au bon moment.

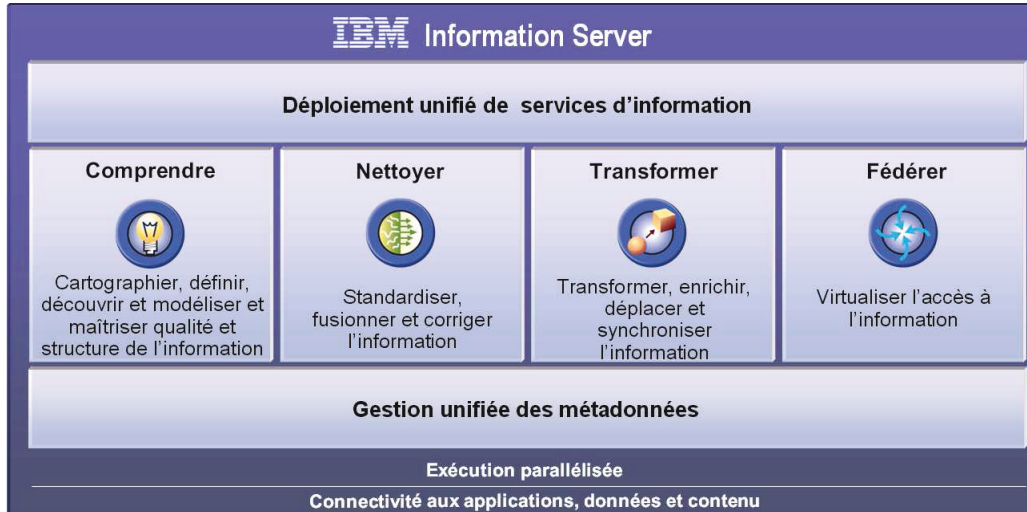


Figure 14 : La suite Information Server

Infosphere Datastage permet d'accomplir les objectifs suivants :

- construire des flots de données qui extraient les informations provenant de sources multiples, transforment les données au format attendu et délivrent les données aux bases de données ou applications cibles.

- se connecter directement aux applications de l'entreprise en tant que sources ou cibles pour s'assurer que les données sont pertinentes, complètes et exactes.
- réduire les temps de développement et améliorer la cohérence de la conception et du déploiement à l'aide de fonctions prédéfinies.
- réduire le cycle de réalisation du projet en travaillant avec un ensemble commun d'outils dans la solution.

Les points forts de l'outil Datastage sont :

- la connectivité à un large éventail de format de fichiers, base de données grâce aux nombreux connecteurs fournis.
- une bibliothèque de plus de 300 fonctions prédéfinies incluant des transformations complexes.
- une exécution des jobs possibles en batch, en temps réel ou comme service Web.
- un traitement parallèle haute-performance.

Datastage est avant tout un outil opérationnel demandant de comprendre la problématique métier avant de se lancer dans l'intégration. Il permet de s'affranchir des problématiques techniques de programmation fastidieuses afin de se concentrer sur la finalité métier. La conception s'approche des méthodologies de développement rapide d'applications grâce à une interface de gestion des flux visuels.

Les conceptions réalisées sont organisées par projet. Ces projets sont des containers qui organisent et fournissent la sécurité aux jobs. Les jobs sont les travaux correspondant à un flux de données au sens large. Il existe deux types de travaux : les travaux parallèles qui sont les travaux unitaires permettant la transformation d'une source en cible et les travaux séquences qui sont les travaux permettant de spécifier une séquence de travaux parallèles à exécuter. Les travaux séquences contrôlent l'interaction et la gestion des erreurs entre les travaux individuels, ils forment ainsi un module de bout-en-bout. Les jobs sont conçus graphiquement et sont composé de stages qui sont les opérateurs élémentaires de traitement des données.

Les principaux stages disponibles pour concevoir un **job parallèle** sont :

- **les stages fichiers**



Sequential File : ce stage permet d'extraire (source) ou de charger (cible) les données dans un fichier texte séquentiel mono-format.



Dataset : ce stage permet de créer ou de lire des fichiers de données dans un format binaire spécifique à Datastage. Les datasets sont utilisés pour transmettre des données entre les jobs Datastage.



Complex Flat File : ce stage permet d'extraire (source) ou de charger (cible) les données dans un fichier texte séquentiel multi-format.



ODBC Connector : ce stage permet d'extraire (source) ou de charger (cible) des données dans une base de données accessibles par des sources ODBC.

- **les stages traitements**



Copy : ce stage permet de copier un flux entrant en n flux sortants.



Aggregator : ce stage permet d'agréger les enregistrements d'un flux à partir d'une clé d'agrégat (équivalent du GROUP BY dans SQL).



Filter : ce stage permet de filtrer les enregistrements d'un flux en paramétrant une condition spécifique sur un ou plusieurs de ces champs (équivalent du WHERE dans SQL).



Join, Merge, Lookup : ces stages permettent de réaliser tous les types de jointures (internes, externes, gauche, droite) possibles entre plusieurs flux entrant. L'utilisation de l'un ou l'autre de ces stages est régit par le volume des flux (notamment la différence entre le flux principal et le flux de référence).



Remove Duplicates : ce stage permet de supprimer les doublons dans les enregistrements d'un flux à partir d'une clé (simple ou composée).



Sort : ce stage permet de trier les enregistrements d'un flux sur une clé (simple ou composée).



Transformer : ce stage est le plus complet, il permet de filtrer les données sur une condition donnée et permet la définition de nouvelles données en réalisant des dérivations basées sur un ensemble de fonctions prédéfinies (fonctions de conversion, de calcul...)

Les principaux stages disponibles pour concevoir un **job séquence** sont :



StartLoop et *EndLoop* : permet de créer des boucles de traitement à partir d'un compteur ou d'une liste d'items.



Job Activity : permet d'exécuter des jobs Datastage (job parallèle ou séquence).



Nested Condition : permet d'intégrer une condition dans le flux (équivalent du IF dans la plupart des langages de programmation).



Notification Activity : permet de déclencher une notification par mail.



WaitForFile Activity : permet de scanner la présence d'un fichier dans un répertoire.



Execute Command : permet d'exécuter une commande sur le système d'exploitation (ligne de commande DOS ou UNIX).

En termes d'architecture logique, la solution est composée des éléments suivants :

- **la plateforme Information Server** (*Information Server Platform*)

Cette partie serveur est constituée :

- du moteur d'exécution parallèle qui exécute les tâches Information Server et qui est constitué de l'Information Server Engine (qui exécute les jobs DataStage), des Service Agents (processus fournissant la communication entre les Services et le Moteur) et des Connectors (qui fournissent l'accès aux données lors du design et de l'exécution).
- du module Services qui est l'ensemble de services partagés permettant de centraliser les tâches de base : les Common Services (qui réalisent les tâches de sécurité, logging, reporting, metadata et exécution), les Product-specific Services (qui fournissent les tâches spécifiques aux produits, DataStage notamment) et le serveur d'application Websphere (moteur de transactions haute-performance permettant de construire et exécuter des applications Web dynamiques impliquant typiquement le protocole HTTP)
- du référentiel (Repository) qui stocke les objets de tous les produits de la suite Information Server. Ce référentiel est une base de données pouvant être hébergé sur de multiples SGBD.

- **les modules clients** (*End Users*)

Trois clients lourds sont fournis pour l'utilisation de Datastage :

- le Designer : qui permet, aux développeurs, la conception graphique des travaux sur les données.
- le Director : qui permet de suivre l'exécution des travaux au cœur des projets Datastage.
- l'Administrator : qui permet de gérer l'ensemble des projets d'un serveur.

La solution est complétée par deux consoles d'administration Web :

- une console pour la gestion du serveur d'application WebSphere permettant notamment de gérer les droits d'accès à la solution.

- une console pour la gestion de la suite InfoSphere offrant notamment la possibilité d'effectuer des exports et imports de composants afin de réaliser les livraisons entre les différents environnements.

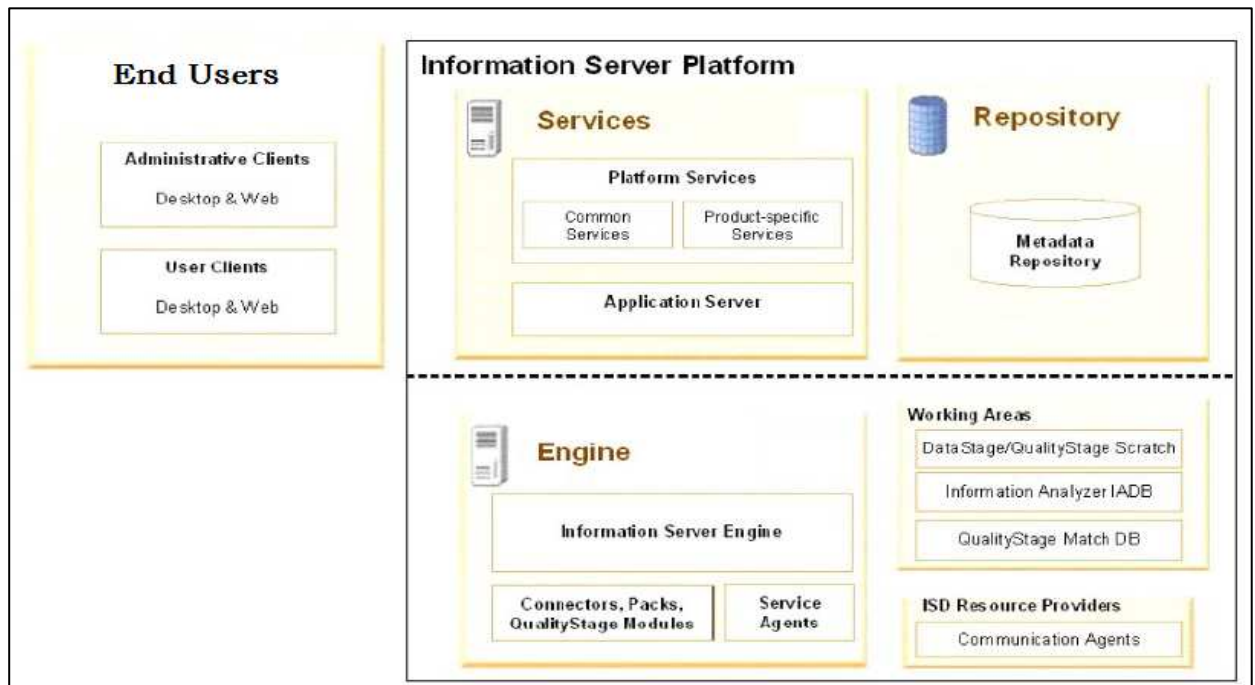


Figure 15 : Architecture logique Information Server

III.4 Le bilan de la phase d'acquisition

La fin de cette phase d'acquisition a été planifiée pour le 31 janvier 2014.

Dans les faits, le choix définitif a bien été réalisé pour cette échéance mais la phase de contractualisation a débordé sur le mois de février.

<i>Tâches</i>	<i>Dates planifiées</i>	<i>Dates réalisées</i>
Rédaction et transmission de l'appel d'offre	Du 01/11/13 au 15/11/13	=
Présentation des solutions	Du 16/11/13 au 29/11/13	=
Réception des réponses des consultés	01/12/13	=
Analyse des réponses	Du 02/12/13 au 19/12/13	=
Sélection des réponses (short-list)	19/12/13	=
Rédaction et transmission d'une expression de besoins	Du 19/12/13 au 02/01/14	=
Soutenances	Du 06/01/14 au 10/01/14	Du 06/01/14 au 15/01/14
Analyse finale des réponses	Du 13/01/14 au 31/01/14	Du 16/01/14 au 31/01/14
Présentation à la Direction Générale	29/01/14	=
Décision finale	31/01/14	=

Tableau 1 : Planning des tâches de la phase d'acquisition

Les principales difficultés rencontrées lors de cette phase d'acquisition ont été :

- exprimer les besoins liés à l'ETL de façon exhaustive mais raisonné.
- comparer les réponses fournies en se basant sur des critères objectifs.
- détacher le discours commercial tenu par les consultés du discours technique.
- expliquer et convaincre les décideurs (direction générale) quant au choix et à l'intérêt de la solution retenue.

IV La mise en place de la solution

La mise en place de l'outil est une phase critique car les performances et donc l'acceptation de la solution découleront directement des choix techniques et de la qualité de l'implémentation. J'ai menée cette phase avec l'aide des équipes infrastructure de la DSI CGA et des intervenants de l'intégrateur de la solution.

IV.1 L'installation de la solution chez CGA

L'architecture retenue pour la solution est constituée d'un environnement de non-production et d'un environnement de production.

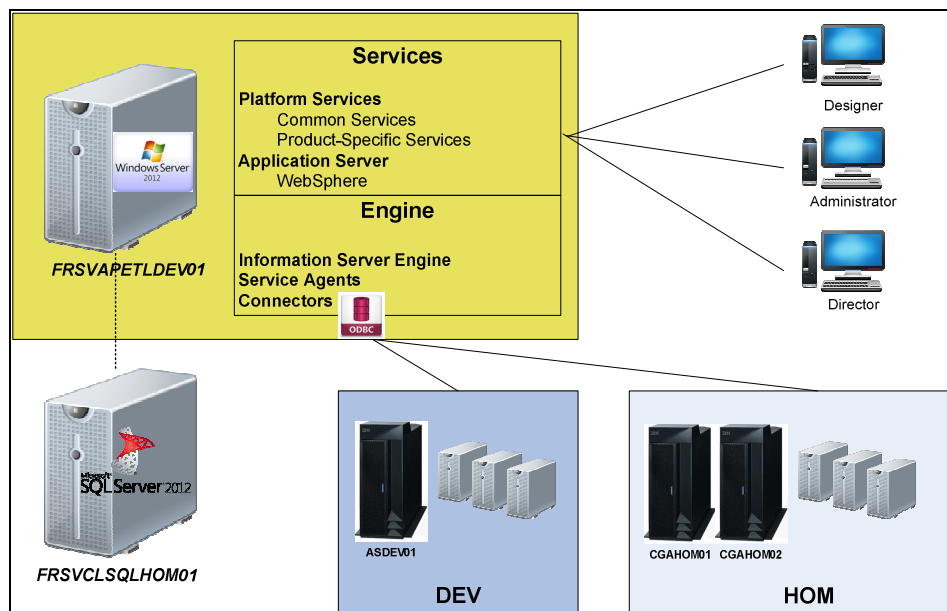


Figure 16 : Environnement de non-production Datastage

L'environnement de non-production est constitué d'un serveur d'application Windows 2008 Server hébergeant le moteur (Information Server) et les services (WebSphere, services communs et services spécifiques Datastage). Ce serveur d'applications accède aux différentes sources et cibles des environnements de développement et d'homologation via des sources ODBC.

Le référentiel de la solution est stocké dans une base SQL Server 2012 hébergée sur un serveur clusterisé identique au serveur de production. Les clients (lourds et légers) permettent de créer les jobs (Designer), de superviser les jobs (Director) et d'administrer les projets (Administrator).

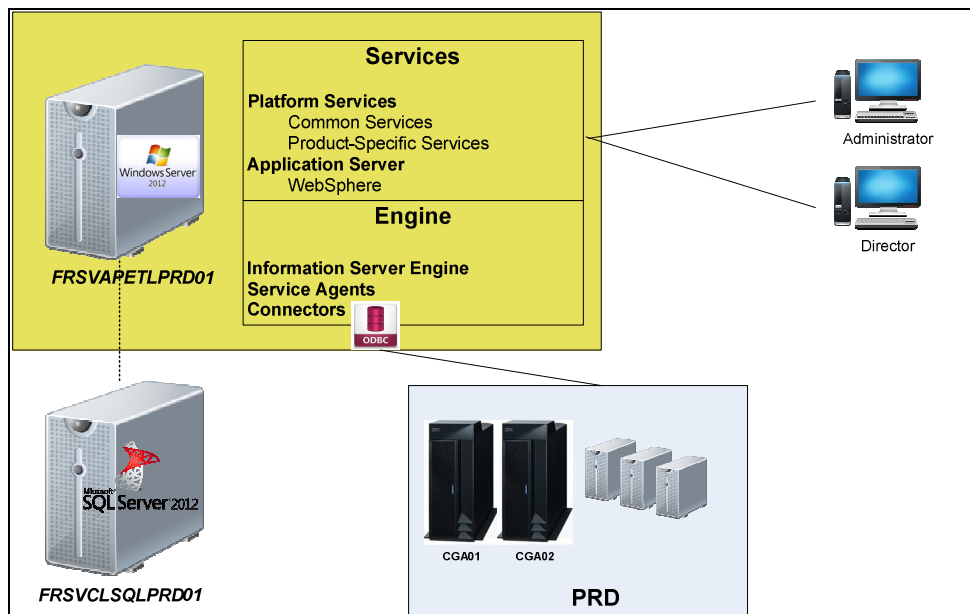


Figure 17 : Environnement de production Datastage

L'environnement de production est constitué d'un serveur d'application Windows Server 2008 hébergeant le moteur (Information Server) et les services (Websphere, services communs et services spécifiques Datastage). Ce serveur d'applications accède aux différentes sources et cibles des environnements de production via des sources ODBC.

Le référentiel de la solution est stocké dans une base SQL Server 2012 hébergée sur un serveur clusterisé. Les clients (lourds et légers) permettent de superviser les jobs (Director) et d'administrer les projets installés (Administrator). Le client permettant de créer des jobs (Designer) ne sera pas installé en production, la création d'un job devant nécessairement faire l'objet d'un développement dans l'environnement de non-production.

J'ai suivi l'installation des deux environnements avec l'intégrateur choisi lors de la phase d'acquisition qui est intervenu dans les locaux de CGA. En parallèle à ces installations, nous avons rédigé une procédure complète d'installation du serveur d'application.

J'ai également rédigé une procédure d'installation des postes clients qui permet à notre service Desk d'être autonome pour le déploiement de la solution sur les postes des personnes concernées.

La phase d'installation de la solution a fait également l'objet d'un dossier d'architecture technique (DAT) permettant de centraliser l'ensemble des informations relatives à l'infrastructure mise en place pour le fonctionnement de la solution Datastage.

J'ai détaillé, avec l'aide du service Infrastructure et Production, dans ce dossier les thématiques suivantes :

- **le schéma d'architecture générale** : celui-ci reprend les figures précédentes (« Environnement de non-production » et « Environnement de production »)
- **les caractéristiques des serveurs** : les deux serveurs d'applications (FRSVAPETLDEV02 et FSVAPETLPRD02) sont des serveurs Windows 2008 R2 virtuels hébergés sur un ESX (chez notre hébergeur pour le serveur de production et au siège pour le serveur de non-production). Le serveur de non-production a une configuration inférieure (2 CPU – 6 Go de RAM – 80 GB de DD) à celui de production (4 CPU – 8 Go de RAM – 80 GB de DD). Ce point est à prendre en compte lors de l'évaluation des performances d'un job ETL avant sa mise en production. La configuration réseau des deux environnements (adresse IP, masque de sous-réseau, DNS...) est également précisée dans le DAT.
- **les composants logiciels pré-requis** : le framework Microsoft .NET 4.0 est requis sur les serveurs ainsi que sur les postes clients. De plus, le serveur de non-production devra être équipé d'un compilateur C++ (Microsoft Visual C++ 2010 Express Edition) pour la compilation des jobs utilisant le stage Transformer.
- **les bases de données dédiées** : chaque environnement dispose d'un référentiel et d'une base de données dédiée à la console d'opération stockés dans une base de données SQL Server 2012 sur un serveur mutualisé. Le DAT décrit ces bases de données (nom, serveur, instance, collation, politique de backup) et les utilisateurs spécifiques créés pour y accéder.
- **les comptes de services créés** : chaque instance nécessite trois comptes de services locaux : un compte administrateur Infosphere (isadmin), un compte administrateur Websphere (wasadmin) et un compte administrateur Datastage (dsadmin).
- **les services créés** : l'installation de la solution génère de nombreux services en raison de ses trois couches (Infosphere, Websphere, Datastage). L'ensemble de ses services sont listés dans le DAT avec leurs caractéristiques (nom, profil

d'exécution, type de démarrage, dépendances). Deux comptes de service sont créés dans le domaine (T-DTSTGE-DEV et T-DTSTGE-PRD) pour exécuter ses services.

- **les ports ouverts pour le bon fonctionnement de la solution** : la politique de sécurité de CGA impose la fermeture de tous les ports sur un serveur. La liste des ports (TCP) dont l'ouverture est strictement nécessaire est donc requise dans le DAT.
- **le plan de sauvegarde** : les besoins en termes de fréquence et de rétention des sauvegardes des serveurs d'application (VM) ainsi que des bases de données sont précisés.
- **le diagramme simplifié des flux** : ce diagramme de flux permet d'appréhender les relations entre le serveur d'applications ETL et les autres serveurs du SI. Le serveur communiquera par ODBC avec les serveurs iSeries (5 partitions actuellement) et les serveurs SQL Server et pourra utiliser les partages réseaux pour échanger avec les serveurs de fichiers Windows.

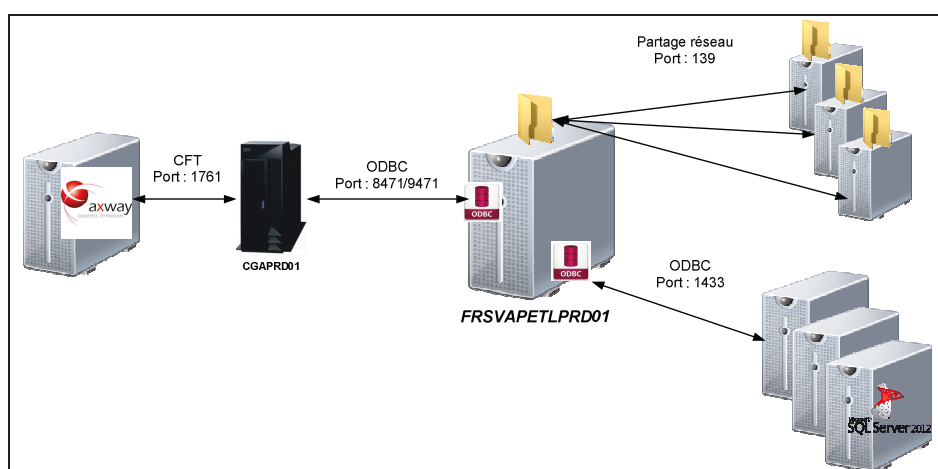


Figure 18 : Diagramme des flux Datastage

En partenariat avec le responsable sécurité, j'ai également renseigné un dossier d'analyse des risques. Ce document détaille les besoins sécurité et les impacts des principales fonctionnalités, en fonction des catégories d'utilisateurs et des types de données utilisées. Il constitue le cahier des charges sécurité de l'application et à ce titre il est l'un des documents de référence du projet. Les fonctionnalités sont évaluées en termes de disponibilité, de preuve et de contrôle. Les données sont évaluées en termes d'intégrité et de confidentialité.

IV.2 L'intégration de l'ETL dans le SI de CGA

L'ETL ayant vocation à devenir l'un des éléments critiques du SI de CGA, j'ai travaillé sur son intégration au sein du SI et sur ses interactions avec les autres outils du SI.

Dans ce cadre, j'ai abordé les thèmes suivants :

IV.2.1 La gestion des droits

La politique de sécurité de CGA impose une authentification aux outils du SI via le login de l'annuaire centralisé (LDAP) ou une authentification SSO (single sign-on) avec l'utilisateur Windows authentifié sur le poste. L'option SSO n'étant pas disponible sur les clients Datastage, j'ai travaillé sur la mise en place de l'authentification LDAP. Celle-ci se réalise en paramétrant le mapping des comptes (utilisateurs et groupes) de l'AD sur la couche WebSphere de la solution. Suite à ce paramétrage, j'ai géré les droits d'accès à la solution Datastage en demandant la création de 6 groupes d'utilisateurs dans l'AD que j'ai pu mapper avec les rôles que je souhaitais utiliser dans la solution.

J'ai ainsi construit les niveaux de droits suivants :

Environnement	Groupe AD	Rôle (Suite)	Rôle (Datastage)
Non-production	Information_Server_Administrators_DEV	Administrator	Administrator
Non-production	DataStage_Administrators_DEV	User	Administrator
Non-production	DataStage_Developers_DEV	User	User
Production	Information_Server_Administrators_PRD	Administrator	Administrator
Production	DataStage_Administrators_PRD	User	Administrator
Production	DataStage_Operators_PRD	User	User

Tableau 2 : Les droits Datastage

Ainsi, il existera des administrateurs complets de la solution pour chaque environnement (groupe « Information_Server_Administrators* »). Les chefs de projets études seront administrateurs de Datastage dans l'environnement de non-production (groupe « DataStage_Administrators_DEV ») afin notamment de pouvoir créer de nouveaux projets. Les développeurs seront simples utilisateurs de la solution dans l'environnement de non-production (groupe « DataStage_Developers_DEV »). Les membres du service exploitation seront administrateur de Datastage dans l'environnement de production

(groupe « DataStage_Developers_PRD ») afin de pouvoir réaliser les mises en production. Les exploitants seront simples utilisateurs de la solution dans l'environnement de production (groupe « DataStage_Operators_PRD ») afin de surveiller l'exécution des jobs et éventuellement de les relancer.

L'accord des droits d'utilisation se fait ainsi simplement en affectant l'utilisateur Windows de la personne concernée au groupe AD détenant les droits nécessaires.

IV.2.2 La mise en production des jobs

CGA dispose avec LMI d'Aldon d'un outil de mise en production automatisé permettant de déclencher des livraisons en ligne de commande. Afin d'utiliser cet outil pour réaliser les mises en production des jobs Datastage, j'ai réalisé un script permettant de prendre en compte les différents scénarios de mise en production.

J'ai identifié deux scénarios différents :

- la mise en production d'un nouveau projet : dans ce cas, la mise en production consiste à créer le projet dans Datastage et à l'initialiser (ceci est rendu possible grâce à la commande en ligne dsadmin) puis à créer l'arborescence externe du projet, importer les variables d'environnement et les composants Datastage (préalablement exportés par le chef de projet initiateur de la demande de mise en production).
- la mise en production d'éléments (nouveaux ou existants) d'un projet existant : dans ce cas, la mise en production consiste uniquement à importer les variables d'environnement et les composants Datastage.

Le script créé (MEP_Datastage.bat) permet de réaliser ces deux scénarios en recevant en paramètres : le nom du serveur d'application Datastage, le nom du projet, un flag permettant de déterminer si il s'agit d'un nouveau projet ou d'un projet existant, le chemin complet du fichier des variables d'environnement à importer, le chemin complet du fichier des composants Datastage à importer et un nom d'utilisateur (et son mot de passe) autorisé à réaliser la livraison.

L'outil d'Aldon permet également de conserver le versioning des composants livrés.

IV.2.3 L'exécution des jobs

CGA dispose d'un scheduler d'entreprise centralisé (ROBOT400) permettant d'exécuter les programmes de manière planifiée ou événementielle. Tous les travaux du SI étant pilotés par ce robot, il est primordial que les jobs Datastage le soient également. J'ai donc réalisé une batterie de tests afin de tester les différentes commandes en ligne et les options fournies par Datastage, j'ai ainsi pu valider la syntaxe que nous utiliserons pour déclencher les jobs est la suivante :

```
dsjob -run -wait -param prs_XXXXX=XXX -param prm_YYYYYY=YYYYYY Projet1 Job1
```

Cette commande permet de déclencher le job *Job1* du projet *Projet1* en lui passant en paramètre les valeurs de *XXX* à l'ensemble de paramètres *prs_XXXXX* et la valeur *YYYYYY* au paramètre *prm_YYYYYY* et attend la fin de l'exécution du job avant de rendre la main au programme appelant.

Datastage possède également un scheduler interne, son utilisation est prohibée au sein de CGA. En revanche, Datastage propose une troisième méthode pour exécuter un job qui est le déclenchement manuel à partir du client Director. Cette méthode pourra être utilisée chez CGA dans le cadre des reprises sur incident réalisées par les opérateurs du service production.

IV.2.4 La supervision des jobs

CGA dispose d'une solution centralisée de supervision pro-active des événements intervenants dans le SI. Pour profiter des fonctionnalités de cet outil, j'ai installé la console d'opérations qui est une option de la solution Datastage. Cette option est constituée d'une console Web permettant de présenter les logs et les statistiques d'exécution des différents jobs Datastage et d'une base de données dédiée (base de données SQL Server 2012 sur le serveur hébergeant le référentiel) permettant de stocker les informations d'exécutions des jobs.

Ainsi, la solution de supervision peut déclencher, à intervalle régulier, une requête simple sur la base de données de la console que j'ai élaboré pour extraire les informations (nom du projet, nom du job, heure d'exécution, statut) des jobs dont l'exécution a échoué et ainsi prévenir par mail les personnes concernées (le service Exploitation).

IV.2.5 Le kit méthodologique

J'ai également travaillé sur l'intégration de l'ETL dans la méthodologie des projets CGA. Pour ce faire, j'ai créé ou complété les documents suivants :

- **modèle de spécifications détaillées**

J'ai créé un modèle de spécifications techniques détaillées permettant aux analystes et aux chefs de projet de décrire précisément les travaux qui devront être réalisés par les développeurs. Ce document est découpé en 3 parties : la description des sources (fichiers, tables), les étapes de transformation et la description des cibles (fichiers, tables).

- **matrice de chiffrage**

Nous utilisons actuellement une matrice nous permettant de réaliser les chiffrages pour nos développements de type Cobol, RPG, WebDev... J'ai complété et amélioré cette matrice pour permettre d'évaluer au mieux les charges induites par le développement d'un nouveau flux ETL. Pour ce faire, je suis parti des hypothèses suivantes, pour l'estimation des charges de réalisation :

- flux simple : création (0.3 J/H) / modification (0.15 J/H)
- flux normal : création (0.6 J/H) / modification (0.3 J/H)
- flux complexe : création (1.3 J/H) / modification (0.6 J/H)
- flux très complexe : création (2 J/H) / modification (1 J/H)

Le modèle applique ensuite les coefficients suivants pour chaque phase de la mise en place du flux.

Description	%
Analyse – Spécifications	17,0%
Réalisation	50,0%
Recette technique	9,0%
Préparation homologation	2,0%
Correction des retours éventuels	10,0%
Préparation mise en production	1,0%
Pilotage (Réunion, Suivi, Documentation)	11,0%
Total	100%

Tableau 3 : Les coefficients appliqués pour le chiffrage des flux Datastage

J'ai pondéré ces hypothèses avec des coefficients de risques liés à l'environnement du projet : maîtrise du sujet par le chef de projet, qualité de l'expression des besoins fournie, expérience de l'équipe chargée de la réalisation, difficultés liées à l'environnement technique, contraintes de sécurité, complexité des opérations d'homologation et de mise en production...

Les coefficients de cette matrice de chiffrage seront affinés en fonction des retours d'expérience sur les futurs développements Datastage.

- **dossier d'architecture**

Chaque projet mené chez CGA fait l'objet d'un dossier d'architecture renseigné par le chef de projet, les architectes et le service infrastructure. J'ai complété ce modèle en intégrant un chapitre « ETL ». Ce chapitre permet notamment de présenter le schéma des flux et de préciser les ouvertures de port nécessaires ainsi que les éventuelles sources de données ODBC supplémentaires à mettre en place sur le serveur d'applications pour accéder aux données sources et cibles.

- **dossier d'exploitation**

Chaque projet mené chez CGA fait l'objet d'un dossier d'exploitation rédigé par le chef de projet à destination du service exploitation. J'ai complété ce modèle en intégrant un chapitre « ETL » qui permet de décrire les flux ETL du projet et leurs problématiques d'exploitation (besoin de sauvegarde, procédure de gestion des rejets / erreurs...).

IV.3 La mise en place des bonnes pratiques d'utilisation de la solution

J'ai tenu à élaborer un document des bonnes pratiques d'utilisation de la solution afin d'optimiser et d'uniformiser l'usage de la solution chez CGA. Ce document est un manuel utilisateur à destination des chefs de projets et développeurs amenés à créer des jobs Datastage et également aux exploitants devant vérifier et reprendre l'exécution des jobs.

Ce document s'inspire des bonnes pratiques recommandées par IBM, de celles formulées par notre intégrateur et de celles partagées par les internautes sur les différents sites que j'ai pu parcourir. J'ai compilé l'ensemble de ces bonnes pratiques en fonction du contexte de CGA et des premiers paramétrages que j'ai pu réaliser. Ce document est amené à évoluer grâce à l'expérience acquise sur l'outil, cependant, afin d'éviter les dérives, ce document est maintenu par un référent qui valide toutes les demandes d'évolution ou de correction.

J'ai organisé ce document sur les thèmes suivants :

IV.3.1 La gestion des projets

Cette section est destinée aux chefs de projet études et décrit l'utilisation des projets (création, modification, suppression). Le projet est l'entité principale de Datastage. Il regroupe des jobs ayant une finalité commune. Une réflexion doit être menée lors de la création d'un nouveau flux pour déterminer si celui-ci doit intégrer un projet existant ou faire l'objet d'un nouveau projet. En effet, un des inconvénients de Datastage est l'impossibilité de partager des jobs entre projets.

Cette section du document décrit les différents paramètres d'un projet et explique notamment comment créer un nouveau projet. Pour ce faire, j'ai créé un script permettant la création automatique des projets. Ainsi, tous les projets de Datastage auront un paramétrage identique.

IV.3.2 La conception des travaux Datastage

Cette section est destinée aux développeurs et décrit les techniques de développement à respecter chez CGA. Je décris dans ce chapitre les règles à suivre pour permettre une

uniformisation maximale des développements afin de faciliter la maintenance dont les conventions de nommage à respecter pour tous les types d'objets (jobs, stages, routines...) et les conventions d'usage des paramètres notamment en mettant en place des ensembles de paramètres avec un jeu de valeurs pour chaque environnement (DEV, HOM et PRD).

J'ai particulièrement travaillé pour élaborer une règle efficace concernant la gestion des environnements AS400, celle-ci étant notre plateforme principale qui contient la majorité de nos fichiers sensibles. Le principe commun sur l'AS400 étant qu'une liste de bibliothèques est toujours définie lors de l'exécution d'un job. Cette liste de bibliothèque provient soit d'une commande passée avant l'exécution du job, soit elle découle des propriétés de l'utilisateur qui déclenche le job.

Exemple :

Pour l'utilisateur User1, la liste de bibliothèque suivante est déclarée : BIB1, BIB2, BIB3.

Pour l'utilisateur User2, la liste de bibliothèque suivante est déclarée : BIB2, BIB3, BIB1.

Le fichier FICA est présent (avec des données différentes) dans les bibliothèques BIB1, BIB2, BIB3.

*Le programme PROGX réalise la requête suivante : SELECT * FROM FICA*

Si l'utilisateur User1 déclenche le programme PROGX les données du fichier FICA de la bibliothèque BIB1 seront extraites.

Si l'utilisateur User2 déclenche le programme PROGX les données du fichier FICA de la bibliothèque BIB2 seront extraites.

Le driver ODBC fournit par la solution Datastage ne permet pas d'utiliser une liste de bibliothèque, la solution est donc de préciser pour chaque fichier le nom de la bibliothèque dans laquelle il est stocké (ex : BIB1.FICA ou BIB2.FICA). Cette solution est peu satisfaisante car elle implique de passer en paramètre les noms de toutes les bibliothèques utilisés dans le job ETL. De plus, on perd la puissance du concept des environnements AS400. Pour pallier à ce problème, j'ai réalisé de nombreux tests qui m'ont amené à privilégier le driver ODBC Client Access qui lui permet de charger la liste de bibliothèque de l'utilisateur connecté.

Dans cette partie, de nombreuses lignes conductrices dans la conception des jobs sont données concernant :

- l'utilisation des fichiers et des bases de données (phase Extract / Load)
- la transformation des données (phase Transform)
- la combinaison des données (phase Transform)
- la restructuration des données (phase Transform)

La gestion des erreurs et rejets fait également l'objet d'un chapitre approfondi afin d'uniformiser les traitements et de simplifier l'exploitation des jobs.

Des techniques de développement avancées permettant d'optimiser les performances sont longuement détaillées. Parmi celles-ci, on trouve l'utilisation du RCP (propagation implicite des champs permettant d'alléger considérablement le travail de conception), l'utilisation optimale des tris (opération la plus couteuse en termes de ressources) et les concepts de partitionnement et de collecte qui sont les bases du parallélisme (ce concept primordiale sera détaillé dans la partie réalisation).

Enfin, les fonctionnalités avancées (débugage, comparaison d'objets, recherche et analyse d'impact, documentation automatique) offertes par la solution Datastage sont expliquées car elles augmentent considérablement l'efficacité dans la conception des jobs.

IV.3.3 L'exploitation de la solution

Cette section est destinée au service Exploitation et décrit les usages liés aux problématiques suivantes : mise à disposition des clients, mise en production des jobs, exécution d'un job, supervision d'un job et gestion des droits d'accès. Ces problématiques sont décrites au chapitre précédent IV-2.

J'ai également animé un débat entre les équipes du service Etudes et du service Production afin de délimiter le périmètre de l'ETL. En effet, la solution Datastage est, de fait, très ouverte car elle permet notamment d'exécuter des commandes (bat, exe, cmd...). Cependant, j'ai souhaité mettre l'accent sur une utilisation des fonctionnalités natives de l'ETL pour lesquelles la solution est très performante mais de ne pas « détourner » l'outil pour des besoins qui ne sont pas dans sa nature.

IV.4 Le bilan de la phase de mise en place

Tâches	Dates planifiées	Dates réalisées
Installation de l'environnement de non-production	28/02/14	31/03/14
Installation de l'environnement de production	30/04/14	25/06/14
Intégration aux outils du SI CGA	31/08/14	=
Rédaction des bonnes pratiques	31/08/14	=

Tableau 4 : Planning des tâches de la phase mise en place de l'outil

Les difficultés rencontrées lors de cette phase de mise en place ont été nombreuses. La principale étant l'impossibilité d'installer l'environnement de production sur un serveur Windows 2012 contrairement aux pré-requis validés par l'intégrateur en début de projet, le comble étant que l'installation avait réussi (après quelques difficultés) dans l'environnement de non-production. Après de multiples interventions de l'intégrateur puis de l'éditeur, la conclusion donnée a été que la version actuelle d'Information Server était supportée mais non certifiée (nuance peu compréhensible) pour Windows Server 2012. Ces difficultés ont été extrêmement chronophages et m'ont amené à faire preuve de diplomatie et de détermination. D'autres problèmes techniques sont intervenus durant cette phase, notamment, l'impossibilité pour Websphere de mapper des noms de groupes de l'AD contenant des espaces, ce qui nous a amené à revoir la convention de nommage en vigueur chez CGA et à remplacer les espaces par des underscores.

L'intégration dans le SI a été également fortement perturbée par les projets en cours et notamment le remplacement du scheduler historique (Robot400) par un nouveau produit (Axway Automator).

Les problèmes rencontrés dans la mise en place de la méthodologie et des bonnes pratiques ont été liés au manque de pratique sur l'outil Datastage mais tous ses aspects seront ensuite affinés grâce à la pratique.

Enfin, j'ai également complété cette phase d'une formation de 4 jours sur l'outil Datastage chez notre intégrateur puis j'ai réalisé le suivi (recueil des besoins, planification) des formations pour les autres membres de la DSI (développeurs et exploitants).

V La réalisation des premiers paramétrages

La phase de paramétrage correspond à la réalisation des premiers flux dans la solution ETL retenue. Cette phase a été découpée en 2 lots dans le cadre de ce projet.

V.1 Le lot 1 : Projets contraints

Ce premier lot est constitué des flux à mettre en place en priorité car intégrés dans des projets contraints par les délais. Ces premiers paramétrages permettront également de se familiariser avec la solution, de compléter le document des bonnes pratiques et d'affiner la matrice de chiffrage des développements.

V.1.1 La migration d'un client majeur sur une nouvelle plateforme de financement

- **Objectif**

Ce premier paramétrage concerne la migration d'un de nos clients principaux d'une plateforme de financement spécifiquement développée chez CGA vers une solution proposée par un éditeur du marché plus performante et plus riche en fonctionnalités. Le client souhaite fortement cette migration mais à condition de ne pas intervenir sur le format des fichiers d'items (factures et avoirs) qu'il nous transmet déjà quotidiennement. Dans ce contexte, l'ETL devra réaliser la conversion du format de fichier existant vers le format attendu par la nouvelle plateforme.

Le principe général est le suivant : notre client nous envoie des fichiers contenant les détails des factures (fournisseur, référence, échéance, montant...) qu'il souhaite mettre à disposition de ses fournisseurs via un portail Web pour que ceux-ci puissent demander un paiement anticipé.

- **Analyse**

Pour ce premier paramétrage, j'ai rédigé un document de spécifications détaillées décrivant les différentes étapes que devra suivre le flux d'items reçus.

Les fichiers transmis par notre client sont au format csv et respectent la structure décrite dans l'annexe 6. Le format attendu par la nouvelle plateforme est également au format csv

avec la structure décrite dans l'annexe 7. On constate que les informations contenues dans les deux formats sont assez proches (au type de données près).

J'ai pu identifier, avec le concours de l'équipe en charge de la nouvelle plateforme de financement, quatre étapes principales dans le flux à mettre en place :

- un aiguillage permettant de router les enregistrements à destination de l'ancienne plateforme en fonction d'un des champs du fichier (le code du donneur d'ordre). En effet, la migration n'est pas totale, certains items doivent encore être déposés sur l'ancienne plateforme.
- des contrôles permettant de rejeter les enregistrements du fichier ne respectant pas les règles d'intégrité : montant négatif pour un même fournisseur et une même échéance, date d'échéance trop ancienne, date d'échéance trop éloignée dans le futur, enregistrement en doublon dans le fichier reçu, enregistrement déjà reçu précédemment.
- une conversion permettant la transformation du format actuel au format attendu par la nouvelle plateforme ainsi que l'enrichissement des données par les données bancaires contenues dans un référentiel.
- un découpage permettant de séparer le fichier reçu en n fichiers. En effet, il s'agit d'une contrainte de la nouvelle plateforme qui n'accepte qu'un fichier par donneur d'ordre.

Les fichiers générés seront déposés dans une bibliothèque de notre plateforme iSeries pour être intégrés dans la nouvelle plateforme.

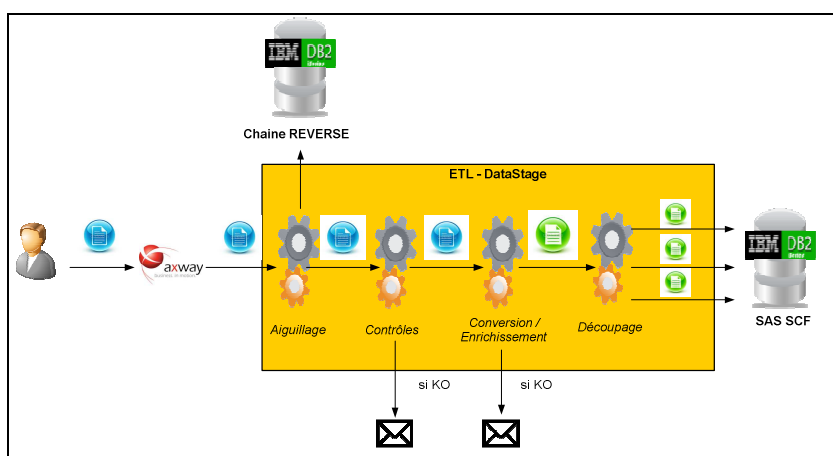


Figure 19 : Analyse du flux Monoprix

- **Réalisation**

Je suis ensuite passé à la phase de réalisation du flux avec l'outil Datastage. J'ai créé le projet « SCF_Monoprix » et le job séquence « SCF_Monoprix_0_Lancement » qui est le job principal qui sera appelé lors de la réception du fichier des items et qui déclenchera les autres jobs parallèles développés.



Figure 20 : Job SCF_Monoprix_0_Lancement (1)

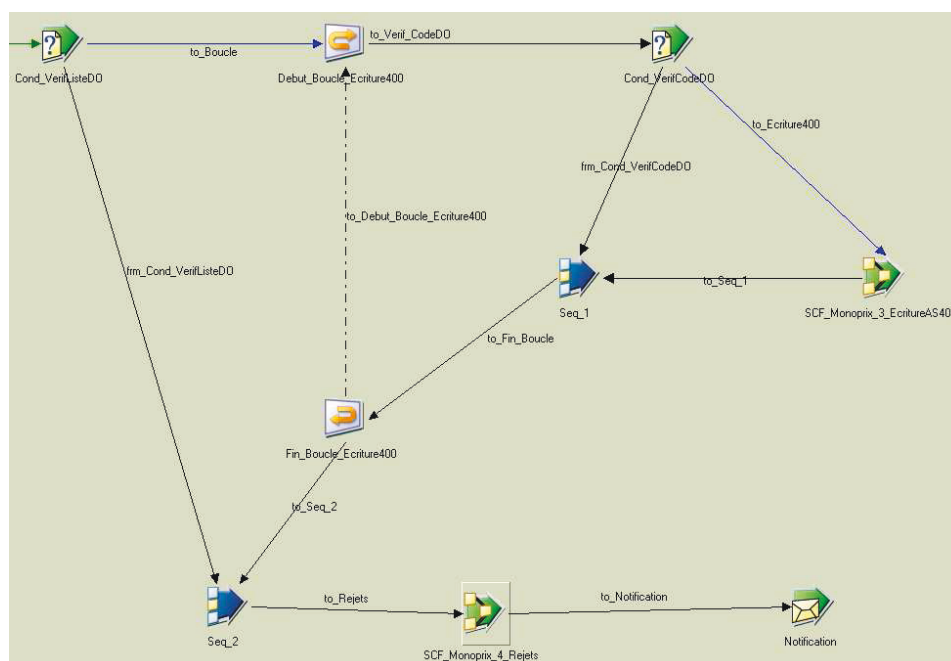


Figure 21 : Job SCF_Monoprix_0_Lancement (2)

Son enchainement est le suivant :

- *SCF_Monoprix_1_Lecture* : appel du job parallèle de lecture du fichier reçu. Ce job lit le fichier reçu sur l'AS400 grâce à un connecteur ODBC (*AS4_FichierIN*) et l'écrit dans un fichier séquentiel déposé en local sur le serveur ETL (*Seq_FichierIN*).

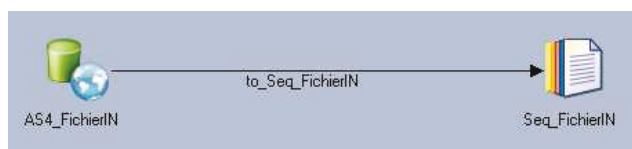


Figure 22 : Job SCF_Monoprix_1_Lecture

- *SCF_Monoprix_2_Conversion* : appel du job parallèle de conversion du fichier reçu. Ce job converti le fichier séquentiel produit par l'étape précédente en un fichier au nouveau format grâce au stage Transformer. Il réalise également les étapes d'aiguillage (container *fn_Aiguillage*), l'enrichissement (grâce au stage Lookup *Lkp_Cpt_Bank* permettant de faire une jointure avec le fichier de référence SCFCPTBNKP présent sur l'AS400) et les différents contrôles (containers *fn_Contrôle**).

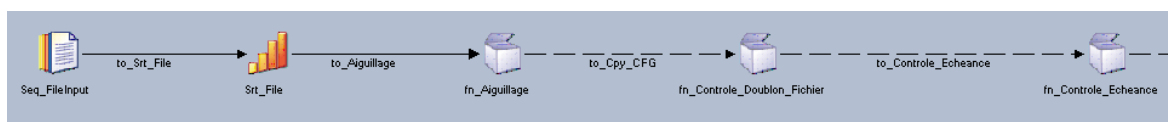


Figure 23 : Job SCF_Monoprix_2_Conversion (1)

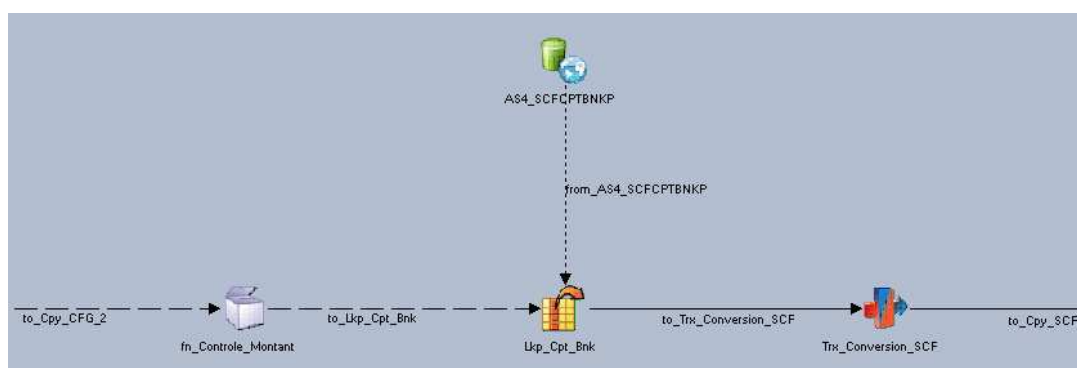


Figure 24 : Job SCF_Monoprix_2_Conversion (2)

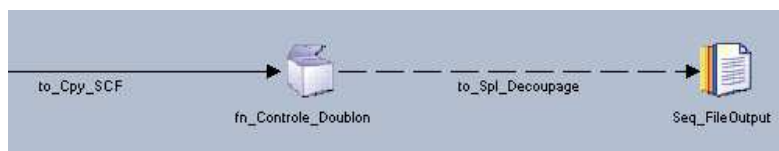


Figure 25 : Job SCF_Monoprix_2_Conversion (3)

Les containers sont des regroupements de stage permettant de simplifier la lecture des jobs et également de regrouper des fonctionnalités réutilisables. J'ai particulièrement travaillé sur les contrôles des données contenues dans les fichiers transmis par notre client. Par exemple, le container *fn_Contrôle_Doublon_Fichier* permet de supprimer les doublons contenus dans le fichier reçu. La logique que j'ai implémentée est la suivante : le flux reçu est copié en 2 flux de sortie identiques grâce à un stage Copy, le second flux est traité par un stage Transformer qui trie le flux selon la clé choisie pour déterminer un doublon (CodeDO – IdFournisseur – DateEcheance – MontantTTC – Sens) et enrichi le flux d'un champ *keyChange* qui vaut 0 si la clé d'un enregistrement est identique à celle de l'enregistrement précédent. Les doublons ainsi identifiés sont transmis vers un fichier

séquentiel permettant de lister les rejets et vers un stage Merge permettant de réaliser une jointure avec le flux initial. Les enregistrements pour lesquels la jointure est un succès sont dirigés vers une impasse, les enregistrements échouant dans la jointure sont ceux n'étant pas en doublons et continuant donc le processus.

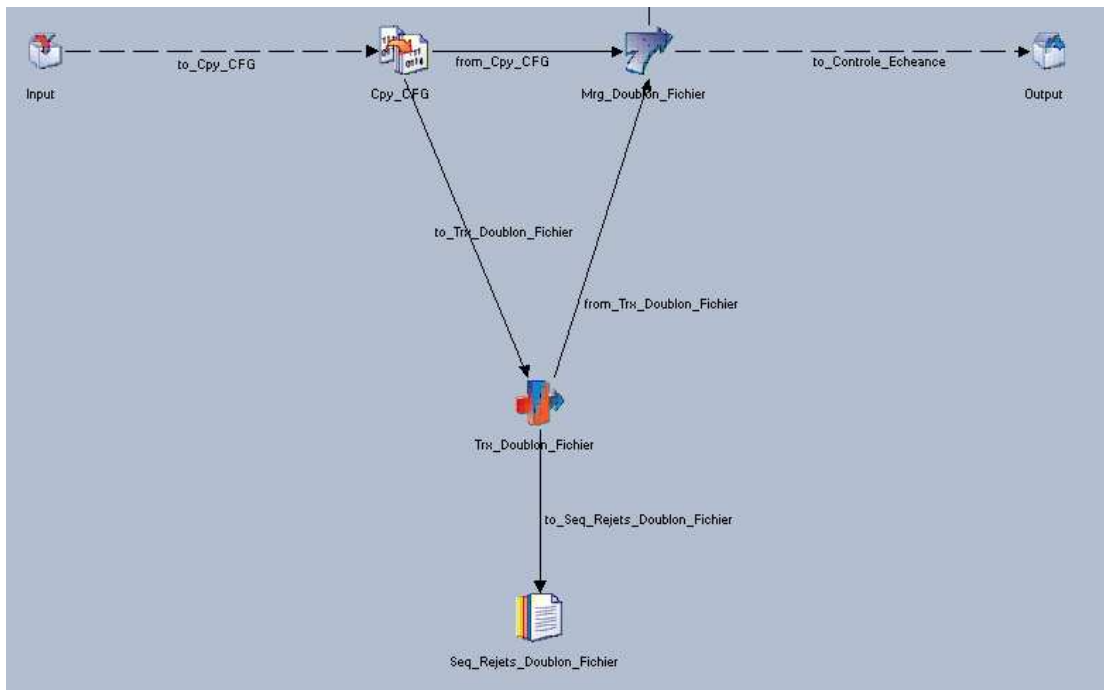


Figure 26 : Container fn_Contrôle_Doublon_Fichier

- *SCF_Monoprix_2_ListeDO* : appel du job parallèle de création de la liste des donneurs d'ordre. Ce job permet de produire un fichier contenant la liste des différents donneurs d'ordre présents dans le fichier de sortie. Pour cela, il élimine les doublons sur le champ CodeDO (équivalent d'un DISTINCT en SQL) puis regroupe les valeurs obtenues en un seul enregistrement séparé par des virgules contenu dans un fichier séquentiel.

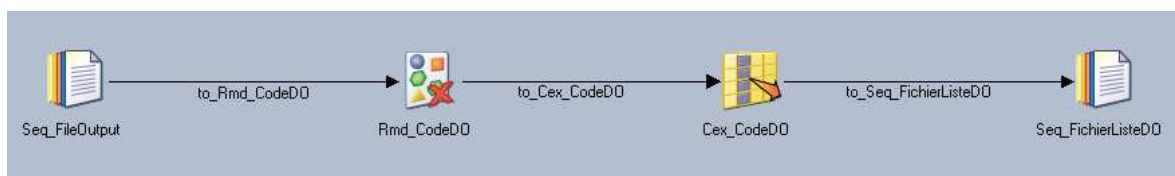


Figure 27 : Job SCF_Monoprix_2_ListeDO

- *Cmd_Liste_DO* : transformation de la liste des donneurs d'ordre en chaîne de caractères. Cette étape est réalisée par l'appel de la commande PowerShell « cat » qui

permet de récupérer le contenu du fichier liste DO produit par l'étape précédente dans une chaîne de caractère.

- *Cond_VerifListeDO* : vérification de la liste des donneurs d'ordre. Cette étape est un test sur la longueur de la liste des DO. Si la liste est vide alors on passe directement à l'appel du job de regroupement des rejets sinon on passe à l'étape suivante.
- *Debut_Boucle_Ecriture400* : boucle sur la liste des donneurs d'ordre. Cette boucle se termine par le stage *Fin_Boucle_Ecriture400*.
- *Cond_VerifCodeDO* : vérification du code DO traité. Ce code ne doit pas être vide pour passer à l'étape suivante.
- *SCF_Monoprix_3_Ecriture400* : appel du job parallèle permettant l'écriture dans un fichier AS400. Ce job filtre les enregistrements du fichier converti sur le code DO qu'il reçoit en paramètre et transforme le contenu du fichier avec un séparateur « ; ».

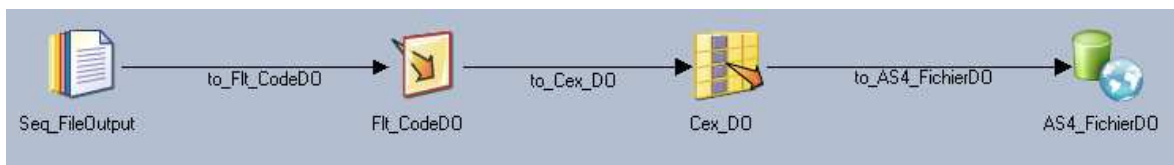


Figure 28 : Job SCF_Monoprix_3_Ecriture400

- *SCF_Monoprix_4_Rejets* : appel du job parallèle de regroupement des rejets. Ce job permet simplement de regrouper les différents fichiers séquentiels de rejets produits par chacun des containers de contrôle dans un fichier de rejets commun.

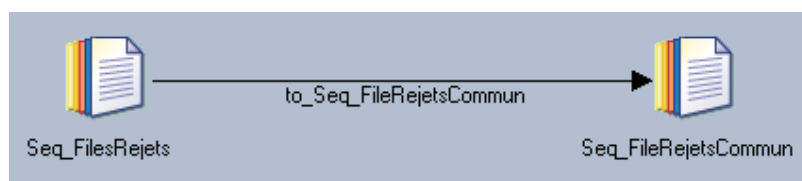


Figure 29 : Job SCF_Monoprix_4_Rejets

- *Notification* : envoi d'un mail aux utilisateurs concernés pour indiquer la réception et le traitement d'un fichier. Le fichier des rejets est attaché à ce mail pour analyse des incohérences.

- **Recette**

J'ai préparé les plans de tests en même temps que la conception. J'en ai profité pour proposer une méthodologie de tests spécifique aux développements ETL car cette phase est particulièrement non mature chez CGA.

- **Livraison / Intégration**

La livraison et l'intégration du flux ont été des étapes chronophages car il s'agissait du premier flux ETL en production chez CGA. J'ai mené ces travaux en impliquant fortement les membres du service Infrastructure et Production afin que ceux-ci arbitrent parmi les directions que j'envisageais.

La solution suivante a donc été validée et mise en place :

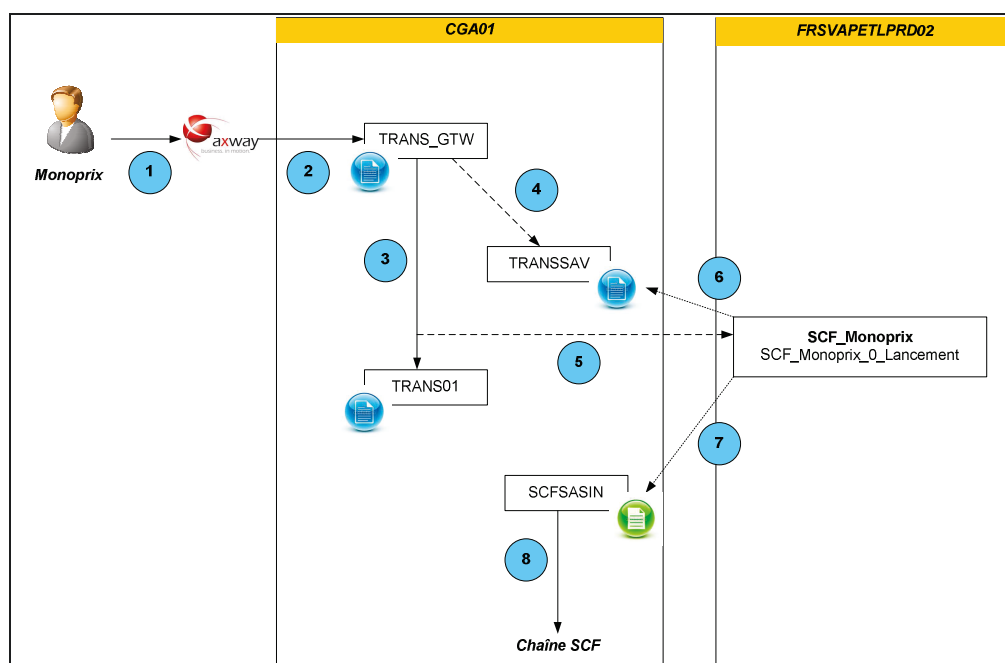


Figure 30 : Schéma de déclenchement du job SCF_Monoprix_0_Lancement

1 : Monoprix continue à nous transmettre ses fichiers d'items par le même flux qu'actuellement.

2 : La Gateway reçoit le fichier provenant de Monoprix et le dépose dans la bibliothèque TRANS_GTW de l'AS400 CGA01. Le fichier est nommé AT*.

3 : BRIDGE400 scan les fichiers AT* présents dans la bibliothèque TRANS_GTW et les transfère dans la bibliothèque TRANS01. La tâche BRIDGE est nommée CGBDMONOP.

4 : La tâche BRIDGE CGBDMONOP réalise également une sauvegarde du fichier dans la bibliothèque TRANSSAV.

5 : Un exit interne sur la tâche BRIDGE CGBDMONOP déclenche le programme SCFMONETLC. Ce programme reçoit en paramètre «FRSVAPETLPRD02 PRD » et déclenche via Robot-Client le job Datastage SCF_Monoprix_0_Lancement en lui passant en paramètre le nom de l'ensemble de paramètres à utiliser (« PRD ») et le nom du fichier de sauvegarde.

6 : Le job Datastage SCF_Monoprix_0_Lancement extrait les données du fichier de sauvegarde.

7 : Le job Datastage SCF_Monoprix_0_Lancement charge les données convertis dans un fichier (ou des fichiers) nommé(s) SCF + CodeDO + ITM et déposé(s) dans la bibliothèque SCFSASIN.

8 : Les fichiers SCF* sont ensuite scannés par BRIDGE400 et entrent dans la chaîne SAS SCF.

J'ai finalement rédigé un dossier d'exploitation décrivant précisément cet enchaînement et les opérations de reprise envisageables en fonction des erreurs rencontrées. J'ai également décrit les traitements à entreprendre en fonction des rejets reçus par mail.

- **Bilan**

Ce premier job Datastage a été livré en production dans les délais imposés par ce projet majeur pour CGA. Deux flux sont traités quotidiennement par ce job. A l'avenir, ce job Datastage devra être intégré dans un projet plus large permettant notamment de mutualiser les contrôles développés pour les besoins d'autres clients.

V.1.2 L'intégration d'une nouvelle filiale

- **Objectif**

CGA intègre une nouvelle filiale en Allemagne. Ce projet est l'un des plus importants de l'année 2014. Dans le cadre de ce projet, l'ETL devra gérer certains flux entre la plateforme Allemande et la plateforme métropole. Le premier flux identifié est celui des données comptables Allemande contenu dans un fichier appelé fichier GL (General Ledger) qui devra être injecté dans le progiciel comptable (COMPTAREL) de la métropole pour être consolidé avec les données comptables de la métropole et transmis à notre partenaire bancaire en Allemagne.

- **Analyse**

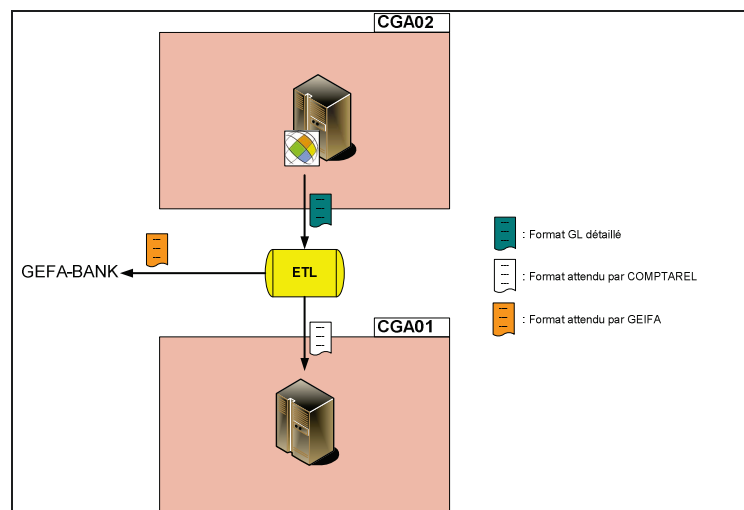


Figure 31 : Analyse de l'intégration du fichier comptable Allemand

Le progiciel cœur de métier installé sur la plateforme iSeries dédiée à la filiale allemande (CGA02) produit quotidiennement un fichier des écritures comptables lors des traitements de fin de journée. Le format de ce fichier est configuré dans le progiciel cœur de métier pour être le plus détaillé possible. L'ETL extrait ce fichier et réalise 2 transformations :

- une transformation permettant de convertir le fichier au format attendu par le progiciel comptable hébergé sur la plateforme iSeries principale (CGA01). Cette transformation consiste principalement à agréger les écritures de même type et à sommer les montants par masse métier.

- une transformation permettant de convertir le fichier au format attendu par le partenaire bancaire Allemand. Cette transformation consiste à produire un reporting en filtrant les écritures pertinentes pour le partenaire.

• Réalisation

Un projet dédié (NAA_FichiersComptables) est créé dans Datastage. Ce projet est constitué d'un job parallèle permettant la transformation du flux en le copiant en deux flux sortants (grâce au stage Copy) puis en agrégeant les données du flux à destination du progiciel Comptable (grâce au stage Aggregator) et en filtrant les enregistrements du flux à destination de notre partenaire bancaire (grâce au stage Transformer).

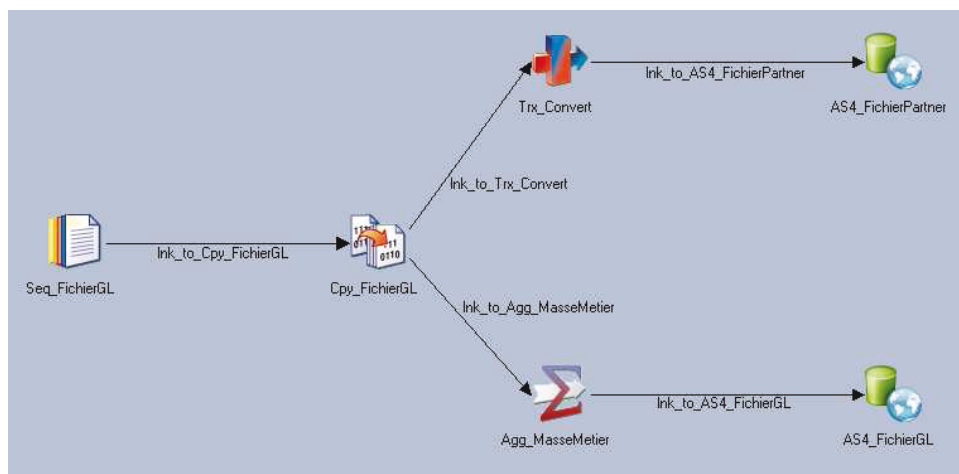


Figure 32 : Job parallèle NAA_FichiersComptables_GL

Le fichier source est un fichier texte présent dans l'IFS (Integrated File System) de la partition AS400 dédiée à la filiale allemande. Ce fichier texte a un nom variable (AQ[timestamp].aq), la propriété *FileName* du connecteur SequentialFile est donc paramétré de manière générique (AQ*.aq). Les fichiers cibles seront écrits sur la partition AS400 de la métropole par l'intermédiaire de connecteur ODBC.

Ce job parallèle assez simple est appelé par un job séquence qui permet d'ouvrir la connexion à l'IFS de la partition allemande (via une commande net use), d'exécuter le job parallèle puis de supprimer le fichier d'origine (via une commande delete) et enfin de fermer la connexion à l'IFS de la partition allemande (via une commande net use /delete).

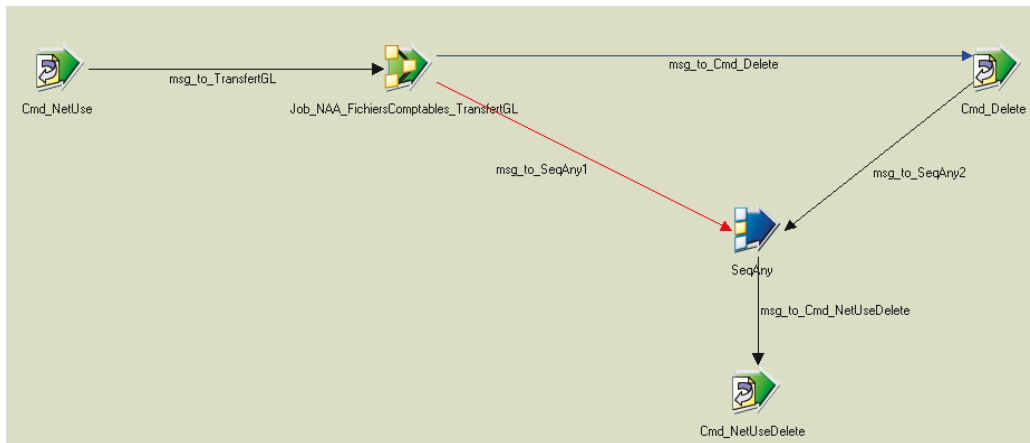


Figure 33 : Job parallèle NAA_FichiersComptables_Lancement

- **Bilan**

Ce paramétrage plutôt simple m’a notamment permis de travailler sur les différentes possibilités de déclenchement d’un job Datastage. En effet, ma première idée était d’utiliser un stage de type *WaitForFileActivity* qui permet de scanner la présence d’un fichier mais ce stage ne permet pas de scanner un nom de fichier générique. J’ai donc abandonné cette option et essayé de créer un job séquence sous forme de boucle avec une temporisation (sous la forme d’une commande système Wait) mais cette solution est peu optimisée. Finalement, j’ai donc opté pour la solution présentée précédemment, avec un nom de fichier générique dans le stage Sequential File et un déclenchement événementiel réalisé par le robot central après le succès du traitement de fin de journée de l’instance allemande du progiciel métier.

V.1.3 La refonte complète du site transactionnel de CGA

- **Objectif**

Un des projets majeurs de CGA sur la période 2014-2015 est la refonte complète du site transactionnel. Dans le cadre de cette refonte, une réflexion est menée quant à la solution idéale permettant les échanges entre le site internet hébergé sur une plateforme Windows avec une base de données SQL Server et notre plateforme principale (iSeries). En effet, de nombreuses communications sont réalisées entre ces 2 plateformes de manière synchrone et asynchrone. Pour la majorité des problématiques, l'utilisation de la solution ETL ne fait aucun doute. Il s'agit :

- de la migration « one-shot » de la version actuelle du site transactionnel vers sa nouvelle version
- du chargement quotidien des données provenant du progiciel cœur de métier vers le site transactionnel
- des transferts asynchrones entre les 2 plateformes

Pour ces différents scénarios, j'interviens en tant qu'architecte quant à la bonne utilisation de l'outil Datastage, le développement étant assuré par une ressource dédiée.

La problématique restante concerne un scénario spécifique du site transactionnel pour lequel nous avons une obligation contractuelle de réponse à l'utilisateur du site en moins de 9 secondes. J'ai réalisé une maquette pour juger de la pertinence de l'utilisation de l'ETL dans le cadre de ce traitement quasi temps réel.

- **Analyse**

Le scénario « Mise à jour du disponible » est le suivant :

- l'internaute fait une demande à partir du site transactionnel pour connaître le montant de son disponible (montant que le client peut se faire financer par CGA).
- sa demande est enregistrée dans une table de la base de données SQL Server du site et un job ETL de déclenchement de la mise à jour est exécuté. Ce job recevra en paramètre un identifiant technique correspondant à la clé primaire de la table SQL Server.

- le job ETL lit l'enregistrement de la table SQL Server et le transforme en fichier XML « question » qu'il dépose dans un répertoire entrée qui est scanné par notre progiciel cœur de métier (Aquarius).
- notre progiciel cœur de métier (Aquarius) traite le fichier XML et répond en générant un nouveau fichier XML de type « réponse » contenant le montant de disponible accordé au client.
- le job ETL lit le fichier XML « réponse » et écrit le résultat dans l'enregistrement de la table SQL Server.

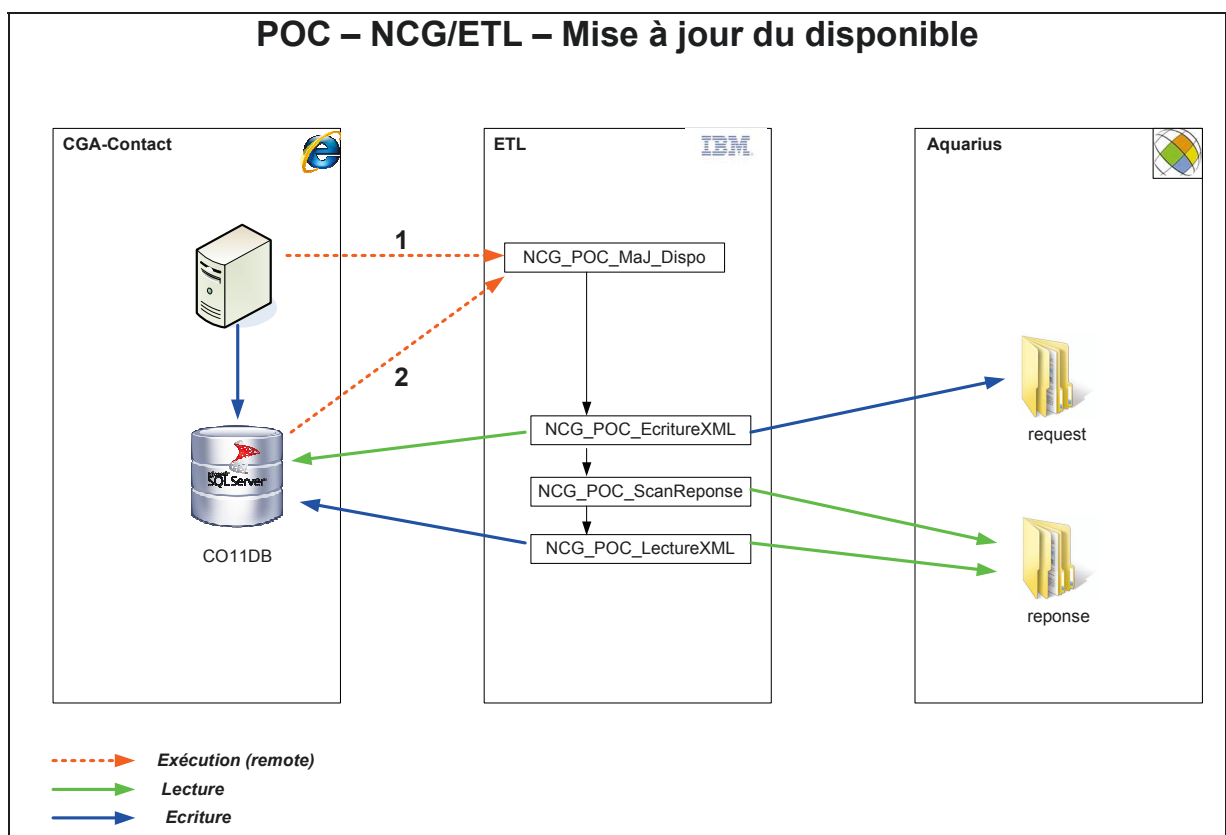


Figure 34 : Analyse de la maquette « Mise à jour du disponible »

• Réalisation du POC (Proof of Concept)

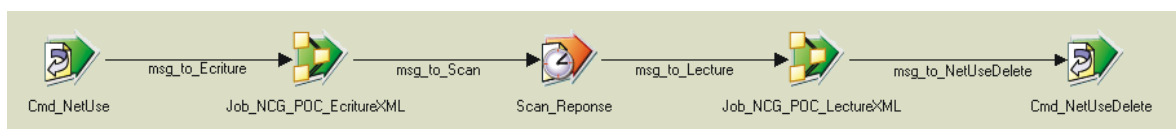


Figure 35 : Job NCG_POC_Maj_Dispo

Un job séquence est créé (*NCG_POC_Maj_Dispo*), celui-ci enchaîne les stages suivants :

- *Cmd_NetUse* : connexion d'un lecteur réseau vers le répertoire d'interface avec le progiciel cœur de métier par l'intermédiaire d'une commande système de type « *net use* ».
- *Job_NCG_POC_EcritureXML* : appel du job parallèle permettant d'extraire l'enregistrement de la table SQL Server du site transactionnel correspondant à l'identifiant reçu en paramètre, de transformer cet enregistrement en un fichier XML et de charger ce fichier XML dans le répertoire scanné par le progiciel cœur de métier. Pour réaliser cette opération, j'utilise le stage « XMLOutput ».

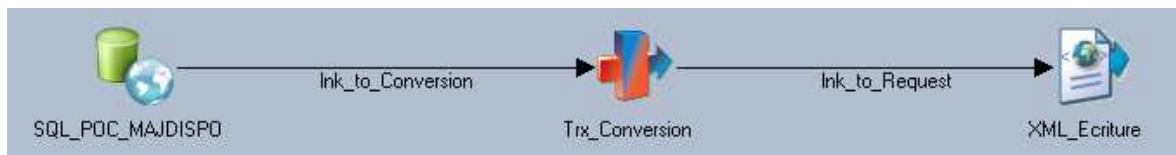


Figure 36 : *Job NCG_POC_EcritureXML*

- *Scan_Reponse* : scan du répertoire dans lequel le progiciel cœur de métier dépose les fichiers XML de réponse.
- *Job_NCG_POC_LectureXML* : appel du job parallèle permettant de transformer le fichier XML réponse et de charger le résultat dans l'enregistrement correspondant à l'identifiant reçu en paramètre dans la table SQL Server du site transactionnel. Pour réaliser cette opération, j'utilise le stage « XMLInput ».

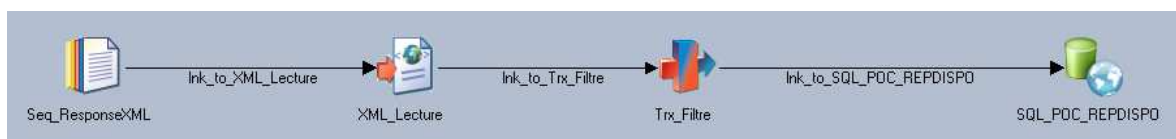


Figure 37 : *Job NCG_POC_Lecture_XML*

- *Cmd_NetUseDelete* : déconnexion du lecteur réseau monté en début de job par l'intermédiaire d'une commande système : « *net use /delete* ».

- **Bilan du POC**

Pour simuler une montée en charge sur le site, j'ai créé un script permettant de déclencher 50 appels du job NCG_POC_Maj_Dispo. Le résultat de cette exécution nous donne un temps de réponse moyen de 18,3 secondes par appel. La contrainte obligatoire des 9 secondes n'est donc pas répondue.

Il existe un temps de traitement incompressible dans l'exécution d'un traitement Datastage (entre 2 et 3 secondes) qui correspond à l'initialisation du job. Ce temps est cumulé à la durée de traitement de l'interface par notre progiciel cœur de métier (2 secondes maximum), au délai du scan (1 seconde maximum), aux délais de connexion aux bases de données et bien sûr aux durées des jobs Datastage eux-mêmes.

La conclusion de ce POC est donc que nous ne pouvons utiliser l'ETL pour les problématiques de « temps réel » ce que nous soupçonnions au départ étant donné la nature d'un ETL. Ces problématiques feront l'objet d'atelier d'architecture dans lesquels nous évoquerons les alternatives possibles et notamment l'utilisation de Web Services.

V.2 Le lot 2 : Rationalisation / Accompagnement

Ce deuxième lot consiste à rationaliser le système d'information en supprimant l'utilisation d'un outil ETL existant chez CGA. Le deuxième objectif de ce lot est d'accompagner la montée en compétence des développeurs en étudiant des problématiques fréquentes dans le traitement des flux chez CGA. L'étude de ces cas particuliers permettra de fournir une méthode appropriée.

V.2.1 Le remplacement d'un outil ETL existant

- **Objectifs**

Un des objectifs majeurs de ce second lot est le remplacement d'un outil ETL (SETI-Map de la société ERT) existant dans le SI de CGA qui nous permettra de rationaliser notre SI ainsi que de réaliser une économie non négligeable au vu des coûts de maintenance de cet outil. SETI-Map est utilisé pour répondre aux besoins suivants : réception de fichier transmis par nos clients par mail sur une adresse générique, transformation éventuelle d'un fichier au format attendu par CGA, dépôt de fichiers dans l'IFS (système de fichiers permettant de partager des fichiers entre le monde Windows et l'AS400) de notre serveur iSeries.

- **Analyse**

Dans un premier temps, j'ai analysé l'ensemble des flux traités par SETI-Map et investigué notamment sur les flux obsolètes. Le résultat de cette analyse préliminaire confirme que 21 adresses mails différentes sont scannées par l'outil, 28 types de flux différents sont traités parmi lesquels un seul fait l'objet d'une conversion. Durant cette analyse, j'ai conçu mes spécifications en pensant à la simplification des flux grâce à la possibilité pour Datastage de travailler directement sur les fichiers AS400 sans passer par l'intermédiaire de l'IFS. J'ai également constaté qu'une amélioration importante pouvait être apportée dans le traitement des mails qui sont aujourd'hui aiguillés sur un seul critère (l'adresse mail de l'expéditeur).

Mon travail d'analyse m'a amené à considérer le principe d'un traitement Datastage qui sera déclenché toutes les 15 minutes par le scheduler principal de CGA. Ce traitement sera découpé en 3 étapes :

- lecture des mails reçus : déclenchement d'un traitement qui permet de lire les mails reçus sur une adresse générique, d'extraire les pièces jointes, de les copier dans un répertoire commun et de créer un fichier d'index de ces mails traités.
- aiguillage des pièces jointes : déclenchement d'un traitement qui permettra d'appliquer les règles de gestion paramétrées pour les pièces jointes reçues à l'étape précédente.
- traitement des fichiers : déclenchement d'un traitement qui permettra d'appliquer les règles de gestion paramétrées pour les fichiers à traiter.

• Réalisation

Datastage ne propose pas de fonctionnalités de traitement de mails en natif. Pour réaliser la lecture des mails reçus j'ai donc développé un programme avec le langage Windev que je ne détaillerais pas ici mais qui permet à chaque exécution de :

- lire les mails reçus sur la boîte aux lettres depuis la dernière exécution.
- copier la ou les pièces jointes de chacun des mails reçus dans un répertoire commun (D:\Projets\ReceptionMail\OPS_TraitementFlux\FICHER\PJ).
- tracer le traitement de chaque pièce jointe dans un fichier (PJ_Recus.csv) au format suivant.

Nom du champ	Type	Description
Destinataire	A(100)	Destinataire du mail
Expediteur	A(100)	Adresse mail de l'expéditeur
Extension	A(3)	Extension de la pièce jointe
Objet	A(255)	Objet du mail
NomPieceJointe	A(255)	Nom de la pièce jointe

Tableau 5 : Format du fichier PJ_Recus.csv

J'ai également créé un fichier de paramétrage (Param_Mails.csv) regroupant les règles de gestion à appliquer sur les différents mails reçus. Ce fichier respecte le format suivant :

Nom du champ	Type	Description
Destinataire	A(100)	Destinataire du mail
Expediteur	A(100)	Adresse mail de l'expéditeur
Extension	A(3)	Extension de la pièce jointe
Objet	A(255)	Objet du mail

PrefixePJ	A(8)	Préfixe de la PJ après renommage
RepertoireDestination	A(255)	Répertoire de dépôt de la pièce jointe

Tableau 6 : Format du fichier Param_Mails.csv

Un job Datastage (TraitementMails_Job) permet de joindre le fichier des mails reçus (PJ_Recus.csv) avec le fichier des règles à appliquer (Param_Mails.csv) et ainsi de produire un fichier de sortie des actions à réaliser (MailsATraiter.csv) au format suivant.

Nom du champ	Type	Description
NomPJ	A(255)	Nom de la pièce jointe
PrefixePJ	A(8)	Préfixe de la PJ après renommage
RepertoireDestination	A(255)	Répertoire de dépôt de la pièce jointe

Tableau 7 : Format du fichierMailsATraiter.csv

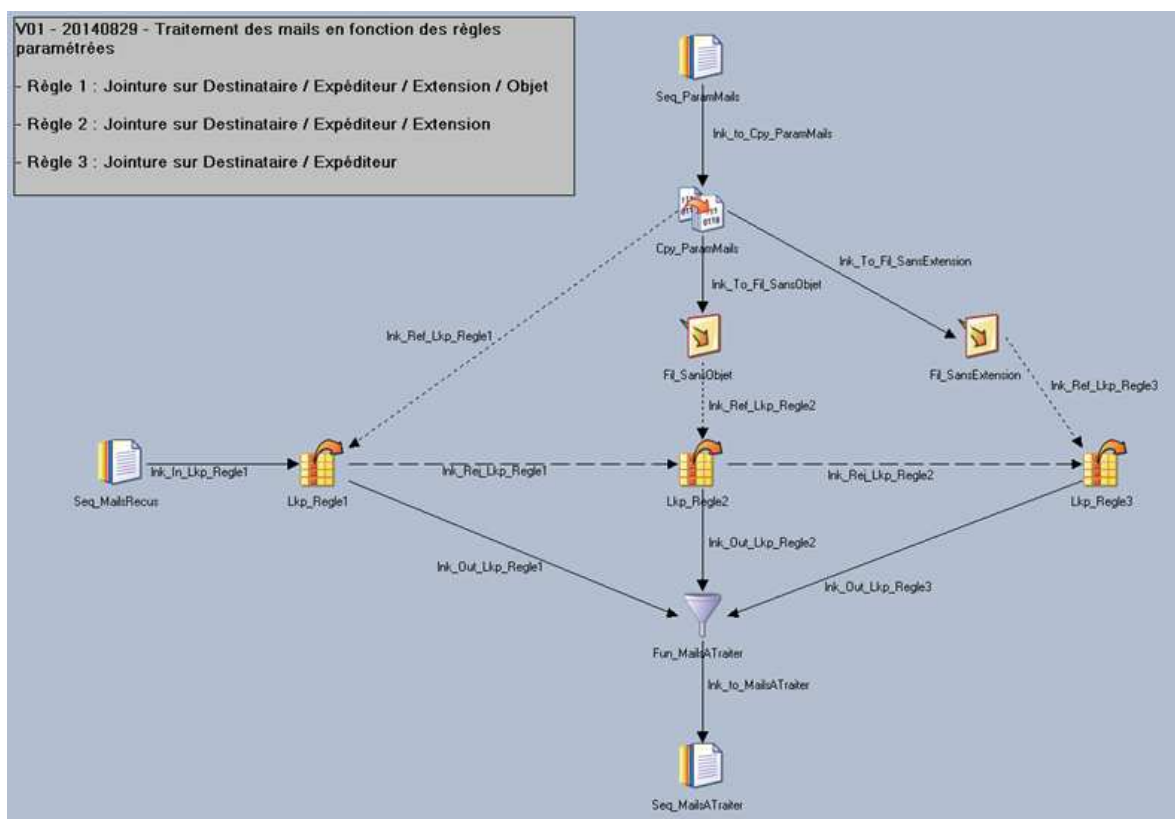


Figure 38 : Job TraitementMails_Job

Le job TraitementMails_Job permet de joindre le fichier des mails reçus avec le fichier des règles paramétrées en 3 jointures successives allant de la règle la plus contraignante à la règle la plus générique. Les enregistrements ne concordant avec aucune règle paramétrée ne seront donc pas intégrés au fichier des mails à traiter.

Le job Datastage suivant (TraitementMails_Seq) réalise une boucle sur le fichier des mails à traiter et pour chacun des enregistrements il réalise le déplacement du fichier vers le répertoire déterminé par la règle de gestion en le renommant avec le préfixe paramétré et un compteur permettant d'éviter les doublons.

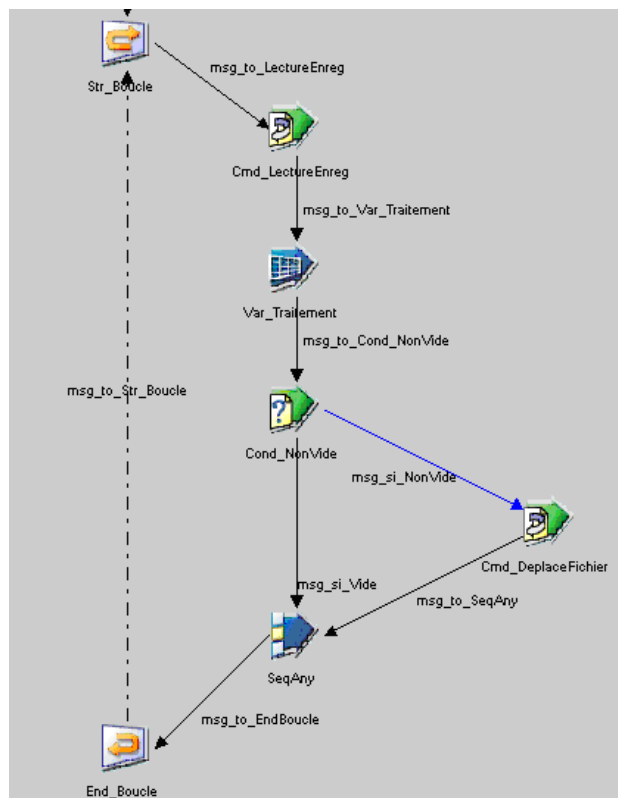


Figure 39 : Job TraitementMails_Seq

Le job Datastage TraitementFichiers_Seq est ensuite déclenché. Celui-ci permet le traitement des fichiers reçus par mail et traités par les étapes précédentes mais également des fichiers qui sont déposés directement dans des répertoires de traitement par les utilisateurs. Les règles de gestion sont paramétrées dans un fichier de paramétrage (Param_Fichiers.csv) au format suivant :

Nom du champ	Type	Description
CheminOrigine	A(200)	Répertoire Windows à scruter
Prefixe	A(8)	Préfixe à scruter
NouveauPrefixe	A(8)	Préfixe du fichier à créer
BibDestination	A(10)	Bibliothèque AS400 de destination
JobDST	A(100)	Nom du job Datastage à créer

Tableau 8 : Format du fichier Param_Fichiers.csv

Le job `TraitementFichiers_Seq` suit la logique suivante :

- boucle sur les enregistrements du fichier des règles de gestion (Param_Fichiers.csv).
- extraction de chacun des champs par l'intermédiaire d'une commande Powershell de type « cat ».
- extraction de l'ensemble des fichiers respectant la règle de gestion traitée.
- boucle sur les fichiers identifiés à l'étape précédente.
- déclenchement du job parallèle paramétré dans la règle de gestion pour traiter le fichier en cours.

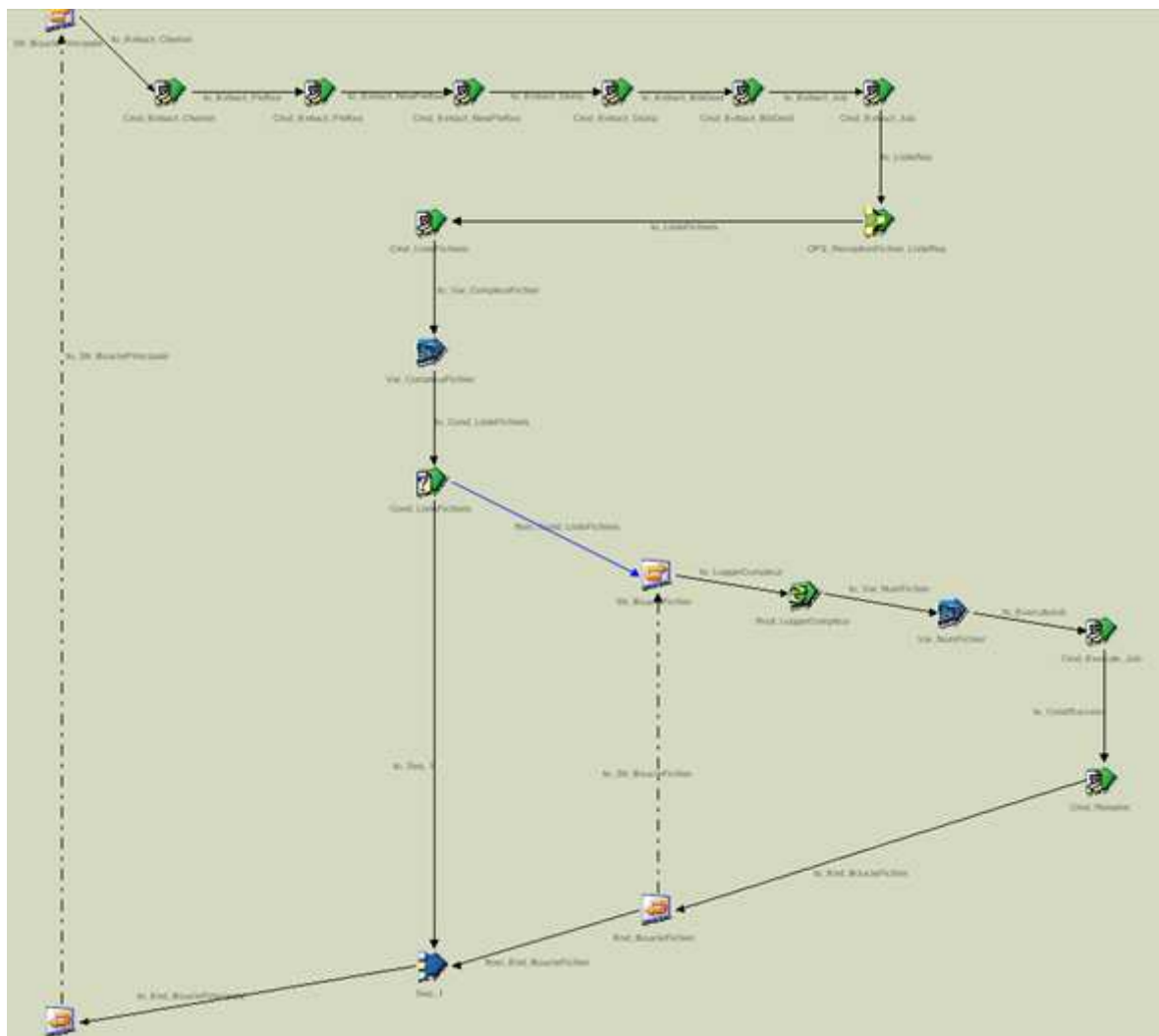


Figure 40 : Job TraitementFichiers_Seq

Le job parallèle paramétré dans les règles de gestion est majoritairement le job GEN_Transfert qui réalise uniquement la conversion d'un fichier CSV de type Windows

en un fichier CSV de type AS400. Cependant, la logique mise en place permet de paramétrer un job Datastage de conversion spécifique pour chaque règle de gestion. C'est actuellement le cas pour un fichier provenant d'un client spécifique pour lequel j'ai créé un job Datastage de conversion du format du client au format pivot CGA.

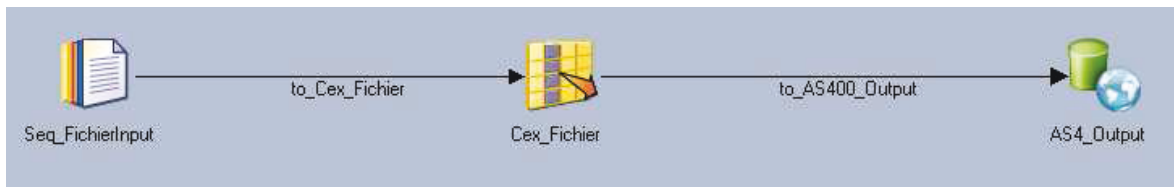


Figure 41 : Job GEN_Transfert

- **Bilan**

Cette solution permet donc une simplification des flux existants actuellement dans SETI-Map et permet également de dé-commissionner cet outil. Le paramétrage des flux existants et futurs est souple. Cependant, la solution n'a pas encore été mise en production car le module de réception des mails fait débat. En effet, le service Infrastructure souhaite éviter la maintenance d'un programme Windev et utiliser les fonctionnalités natives de l'outil Gateway d'Axway qui permet de traiter les mails et d'en extraire les pièces jointes. Cette solution est actuellement à l'étude mais celle-ci ne remet pas en cause les étapes suivantes du projet Datastage.

Un deuxième lot sur cette solution sera envisagée avec notamment la possibilité de déplacer les fichiers de paramétrage actuellement en CSV dans des tables SQL Server. Ce lot devra également prendre en compte les problématiques de purge non traitées dans cette réalisation.

V.2.2 L'étude pratique sur les fichiers multi-format

- **Objectif**

Un nouveau client souhaite nous transmettre directement ces fichiers de remise de factures produits par son ERP (SAP). Ces fichiers ont la particularité d'être multi-format. L'objectif est donc de transformer ces fichiers multi-format au format pivot mono-format utilisé par nos applications. Ce cas pratique pourra inspirer de nombreux autres développements de traitement des fichiers multi-formats notamment pour les fichiers bancaires de type relevé de compte qui sont généralement constitués d'enregistrements de type « en-tête », d'enregistrements de type « détail » et d'enregistrements de type « pied ».

- **Analyse**

Le client nous transmet un fichier multi-format avec séparateurs tabulation constitué de 4 types d'enregistrements :

- type « Header » : enregistrement d'en-tête de fichier contenant notamment l'identifiant du client et la date de création du fichier. Chaque fichier contient un enregistrement de type « Header » qui est identifié par le segment « #H# » en début d'enregistrement.
- type « Payment » : enregistrement de paiement contenant les données principales d'une échéance : fournisseur, montant total, devise, date d'échéance... Chaque fichier contient de 1 à n enregistrement(s) de type « Payment » qui sont identifiés par le segment « #T# » en début d'enregistrement.
- type « Reference » : enregistrement de référence contenant les informations relatives à un item particulier : référence, date du document, montant. Chaque fichier contient un nombre illimité d'enregistrements de type « Reference » chacun étant lié à un enregistrement de type « Payment ». Ce type est identifié par le segment « #R# » en début d'enregistrement.
- type « Footer » : enregistrement de fin de fichier récapitulant notamment le nombre d'items contenus dans le fichier et le montant total. Chaque fichier contient un enregistrement de type « Footer » qui est identifié par le segment « #F# » en début d'enregistrement.

Les formats de ce fichier sont décrits dans l'annexe 8. Ce fichier doit être converti au format pivot de CGA décrit en annexe 7. La conversion doit donc produire un fichier d'items contenant un nombre d'enregistrements égal au nombre d'enregistrements de type « Reference » dans le fichier source. Chacun des enregistrements de type « Reference » étant enrichi par les informations de l'enregistrement de type « Payment » lié.

- **Réalisation**

J'ai testé, dans un premier temps, le stage « Complex Flat File » de Datastage qui est destiné au traitement des fichiers multi-format. Ce stage ne m'a pas donné satisfaction en raison des difficultés de paramétrage et de maintenance. L'intégrateur de la solution m'a également déconseillé son utilisation.

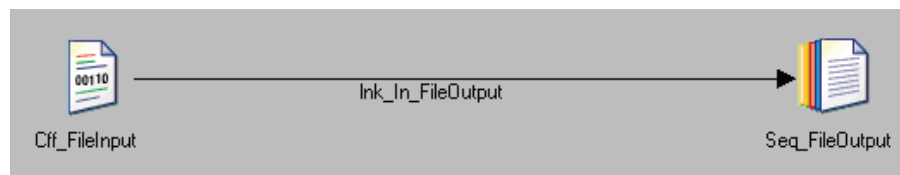


Figure 42 : Utilisation du stage «Complex Flat File»

J'ai donc étudié une solution basée sur un stage « Transformer » pour résoudre les problématiques de multi-format. La logique que j'ai élaborée est la suivante :

- un fichier séquentiel contenant le fichier d'entrée avec un unique champ de longueur 1000 contenant donc chaque enregistrement complet.
- un stage de type « Transformer » : ce stage permet de séparer le flux entrant en 2 flux sortants distincts. Le premier flux sortant contient les enregistrements de type « Payment » identifié par le segment « #T# ». Ces enregistrements sont découpés au format « Payment » grâce au séparateur tabulation et enrichis d'un identifiant technique GroupID. Le deuxième flux sortant contient les enregistrements de type « Reference » identifié par le segment « #R# ». Ces enregistrements sont découpés grâce à des fonctions de sous-chaine pour séparer les informations de référence, date et montant. Ces enregistrements sont également enrichis d'un identifiant technique GroupID et d'un identifiant technique GroupIDParent qui permettra de faire le lien avec l'enregistrement « Payment » correspondant.

La condition nécessaire au bon fonctionnement de cette logique est que le Transformer doit être paramétré en mode séquentiel et non parallèle pour pouvoir traiter les enregistrements dans l'ordre du flux source.

- un stage de type « Jointure » : cette étape permet de joindre grâce à la clé technique GroupIDParent les 2 flux précédemment produit par le Transformer puis de trier le flux résultant sur la clé GroupID.
- un stage de type « SequentialFile » pour charger le résultat dans un fichier CSV (séparateur « ; ») au format pivot connu chez CGA.

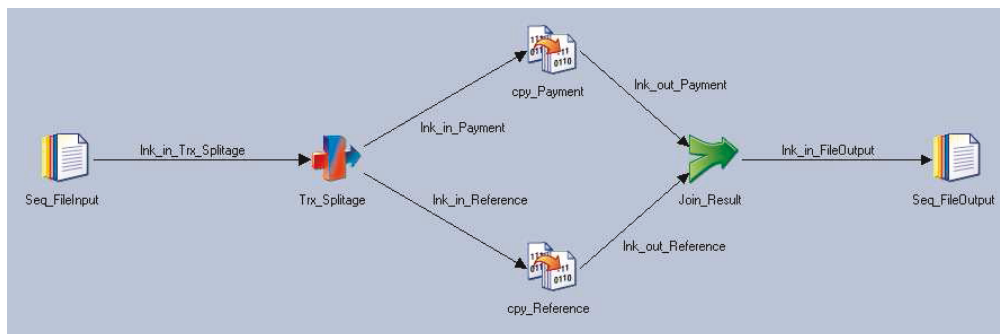


Figure 43 : Job SCF_CoverseionMultiFormat

• Bilan

Ce projet Datastage n'a pas été mis en production immédiatement en raison de retard pris dans la relation avec le client. En revanche, j'en ai préparé une standardisation permettant son utilisation généralisé pour de potentiels clients utilisant des ERP produisant des fichiers multi-formats et notamment SAP.

V.2.3 L'étude sur l'intégration des fichiers volumineux

- **Objectif**

L'objectif initial était de travailler sur l'intégration d'un fichier très volumineux (9 millions d'enregistrements) que nous transmet mensuellement l'INSEE et qui contient l'intégralité des entreprises françaises enregistrées (avec leur raison sociale, leur SIREN, leur type d'activité...). Ce fichier est transformé pour être intégré dans nos différents référentiels sur des plateformes distinctes (DB2/400, SQL Server et Oracle). Les délais impartis étant trop court, je n'ai pu aller au bout de cet objectif mais j'ai réalisé une étude du concept de parallélisme qui est le concept principal d'optimisation des performances dans l'outil Datastage. Cette étude permettra de donner une méthodologie à suivre pour les nombreux développements impliquant des volumes conséquents (chargement du Datawarehouse, chargement du site transactionnel...).

- **Etude du parallélisme**

Le parallélisme partitionné est la clé de la scalabilité des jobs parallèle de Datastage. Le parallélisme est extrêmement performant grâce à la gestion du partitionnement des données réalisée par le framework Orchestrate.

Le partitionnement consiste à séparer les données pour les traiter en parallèle dans des unités de traitement différentes ce qui permet d'améliorer considérablement les performances en optimisant l'utilisation des ressources machines.

Le partitionnement est un élément prépondérant dans la conception des flux. Datastage propose un mode « Auto-Partitioning » qui permet aux développeurs débutants de construire des flux de données sans comprendre en détails les principes du parallélisme. En général, l'auto-partitionnement assure des résultats corrects quand il est utilisé avec des stages standards. Cependant, cette méthode n'est pas toujours la plus efficace et dans certains cas peut amener à des résultats dégradés. Les différents modes de partitionnement et de collecte doivent donc être compris par le développeur pour éviter les mauvaises surprises en termes de performance.

Datastage utilise un fichier de configuration (APT_CONFIG_FILE) afin de gérer les traitements parallèles et les répartir sur les différentes partitions paramétrées. Ce fichier

décrit les différents nœuds utilisés sur le serveur pour le traitement des processus générés par les jobs. Chez CGA, nous utiliserons 4 nœuds différents pour la parallélisation des processus.

Lors de la conception, le développeur pourra utiliser les « partitionneurs » qui distribuent les enregistrements d'un lien dans des segments plus petits qui peuvent être exécutés indépendamment en parallèle. Il existe 2 méthodes de partitionnement :

- le partitionnement sans clé : il distribue les enregistrements sans regarder les valeurs actuelles des données. Les méthodes de partitionnement sans clé sont les suivantes :
 - o Same : conserve le partitionnement existant du stage précédent.
 - o Round-Robin : distribue les enregistrements équitablement entre les partitions (similaire à une distribution de cartes).
 - o Random : distribue les enregistrements équitablement entre les partitions. La méthode Random est moins utilisée que la méthode Round-Robin car moins efficace.
 - o Entire : distribue la totalité des enregistrements à chaque partition.
- le partitionnement avec clé : il examine les valeurs des données dans une ou plusieurs colonnes clés et s'assure que les enregistrements avec les mêmes valeurs dans ces clés sont assignés à la même partition. Le partitionnement avec clé est utilisé quand les règles métier ou les besoins d'un stage requiert l'exécution sur un groupe d'enregistrement liés. Les méthodes de partitionnement avec clé sont les suivantes :
 - o Hash : affecte les enregistrements avec les mêmes valeurs clés à la même partition en utilisant un algorithme de hachage interne.
 - o Modulus : affecte les enregistrements avec les mêmes valeurs clés (pour une clé de type entier) à la même partition en utilisant un simple calcul de modulo.

Le développeur devra également maîtriser les méthodes de collecte qui permettent de combiner les partitions parallèles d'un lien avant un traitement séquentiel. Les méthodes de collecte proposée par Datastage sont :

- Auto : les enregistrements sont lus depuis les partitions sources sans blocage si un enregistrement est indisponible dans une partition particulière. Pour cette raison,

l'ordre des enregistrements dans un collecteur Auto est indéfini et peut varier entre 2 exécutions d'un job sur les mêmes données. Auto est la méthode de collecte par défaut.

- Round-Robin : les enregistrements sont lus depuis les partitions sources en les lisant dans un ordre round-robin. La méthode Round-Robin est généralement plus lente que la méthode Auto parce qu'elle doit attendre qu'un enregistrement apparaisse dans une partition particulière.
- Ordered : les enregistrements de la première partition sont lus puis les enregistrements de la partition suivante, jusqu'à ce que tous les enregistrements soient collectés.
- Sort Merge : si les enregistrements sources sont triés en parallèle, le collecteur Sort Merge génère un flux séquentiel d'enregistrements globalement trié.

- **Méthodologies retenues**

Cette étude m'a permis de fournir aux développeurs une méthodologie permettant de choisir les méthodes de partitionnement et de collecte appropriées.

J'ai élaboré la méthodologie de partitionnement en me basant sur les axiomes suivants :

- la méthode de partitionnement doit donner un nombre de lignes équitables dans chaque partition pour minimiser les surcharges.
- la méthode de partitionnement doit répondre aux besoins métiers et aux exigences fonctionnelles.
- la méthode de partitionnement ne doit pas être trop complexe.

J'en ai déduit la méthodologie de partitionnement suivante :

- commencer par utiliser la valeur par défaut (Auto-Partitioning).
- choisir la méthode Hash pour les stages qui requiert des groupes d'enregistrement liés (Join, Aggregator...).
- choisir la méthode Round-Robin pour distribuer les données équitablement entre les partitions.
- utiliser la méthode SAME pour optimiser le partitionnement bout-en-bout et minimiser le repartitionnement.

J'ai également pu déterminer la méthodologie de collecte suivante :

- commencer par utiliser la valeur par défaut (Auto) si l'ordre de sortie n'a pas d'importance.
- choisir la méthode Sort Merge pour produire un simple flux d'enregistrements triés (si les enregistrements sources ont été triés en parallèle).
- choisir la méthode Round-Robin pour reconstruire les enregistrements dans l'ordre du flux source si celui-ci a été partitionné en mode Round-Robin.

V.2.4 La mise en place de templates

- **Objectif**

Le but de cette tâche est de créer des templates pour les types de flux les plus couramment utilisés chez CGA.

- **Analyse**

Un projet spécifique DST_Templates est créé. Ce projet contient l'ensemble des templates maintenus par le référent Datastage chez CGA. Ces templates sont des jobs qui peuvent être exportés puis importés dans les projets qui souhaitent les utiliser. Pour initier ce projet, j'ai créé des flux simples constitués des 3 phases principales (Extract – Transform – Load) et basés sur les différentes combinaisons possibles entre les types de sources et cibles existantes dans le SI de CGA.

- **Réalisation**

- le job *DST_Template_CSV_DB2* consiste à extraire les données d'un fichier CSV, transformer ces données puis les charger dans un fichier DB2/400 sur l'AS400.

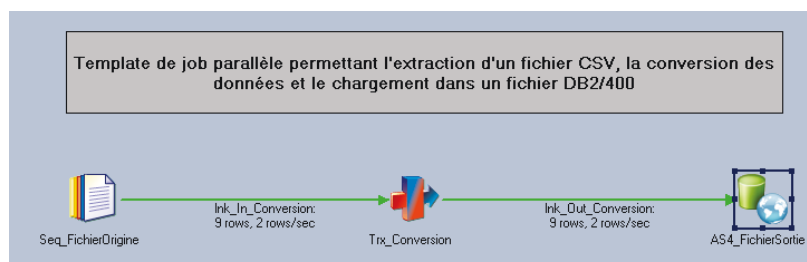


Figure 44 : Job *DST_Template_CSV_DB2*

- le job *DST_Template_CSV_Plat400* consiste à extraire les données d'un fichier CSV, transformer ces données puis les charger dans un fichier plat sur l'AS400. La différence avec le template précédent concerne le fichier cible qui n'est pas un fichier colonné mais un fichier constitué d'un unique champ de type « char » dans lequel les données sont séparés par des « ; ». Pour réaliser cette transformation, on utilise un

stage de type ColumnExport qui permet de concaténer la valeur des différents champs et de déterminer un séparateur.

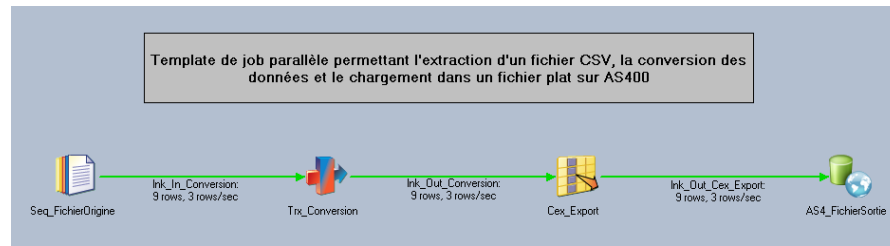


Figure 45 : Job DST_Template_CSV_Plat400

- le job *DST_Template_DB2_CSV* consiste à extraire les données d'un fichier DB2/400, à les transformer puis à les charger dans un fichier CSV.

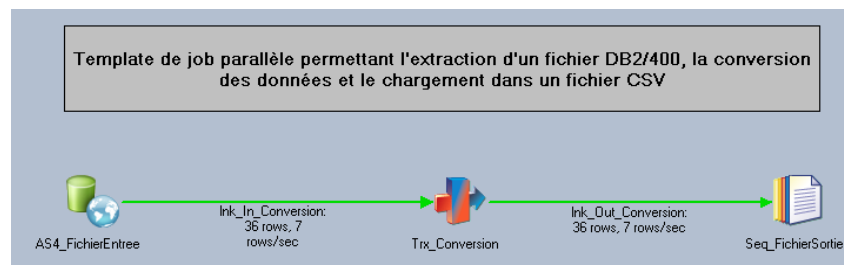


Figure 46 : Job DST_Template_DB2_CSV

- le job *DST_Template_DB2_SqlServer* consiste à extraire les données d'un fichier DB2/400, à les transformer puis à les charger dans une table SQL Server.

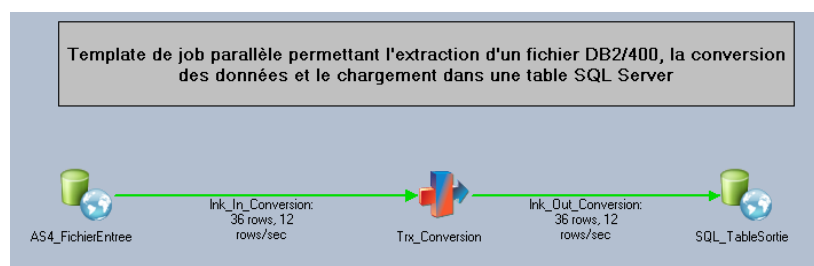


Figure 47 : Job DST_Template_DB2_SqlServer

- le job *DST_Template_SqlServer_CSV* consiste à extraire les données d'une table SQL Server, à les transformer puis à les charger dans un fichier CSV.

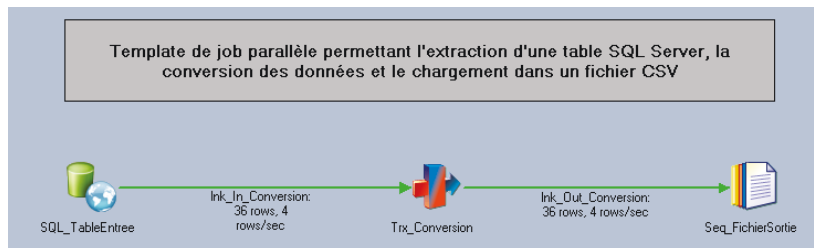


Figure 48 : Job DST_Template_SqlServer_CSV

- **Bilan**

Ces templates peuvent être utilisés en leur passant en paramètres les variables relatives à l'environnement : nom de fichier pour un fichier CSV ou un fichier AS400, nom de la table SQL Server, répertoire de dépôt pour un fichier CSV, nom de la bibliothèque AS400, données de connexion à l'AS400, données de connexion au serveur SQL...

Pour utiliser ces templates, il convient également de charger les schémas des sources et des cibles puis de définir les transformations spécifiques au flux développé.

Cette solution est satisfaisante pour ces templates simples et qui ne sont pas amenés à évoluer. En revanche, je prévoyais de créer également dans ce projet des jobs d'une plus grande complexité dans le but de pouvoir les partager entre les différents projets développés chez CGA mais le fonctionnement de Datastage ne permet pas de partager des objets entre projets. La seule solution envisageable dans ce cadre, est d'exporter le job commun que l'on souhaite utiliser à partir du projet DST_Templates et de l'importer dans le projet en cours de réalisation. Cette solution est peu satisfaisante car elle implique une grande complexité dans la maintenance et la mise à jour des versions. Nous fonctionnerons donc seulement avec des objets partagés (jobs, routines...) au sein d'un même projet en utilisant notamment les possibilités fournies par les containers partagés.

V.3 Le bilan de la phase réalisation

<i>Tâches</i>	<i>Dates planifiées</i>	<i>Dates réalisées</i>
Lot 1 : Projets contraints	31/05/14	15/06/14
Lot 2 : Rationalisation / Accompagnement	10/08/14	15/09/14

Tableau 9 : Planning des tâches de la phase réalisation

Les tâches identifiées en début de projet dans le lot 1 ont été réalisées avec deux semaines de retard sans incidences. Ce délai était contraint par les plannings des projets concernés. Le lot 2 a été prolongé jusqu'au 15 septembre. Sa date butoir était déterminée par des contraintes d'organisation relative à mon changement de poste. Le périmètre de ce lot 2 a fortement évolué en cours de projet.

L'objectif de cette phase a été atteint avec la mise en production des premiers flux et l'accompagnement de la montée en puissance des développeurs en leur fournissant des méthodes et des exemples concrets.

La principale difficulté rencontrée a été la charge de développement qui avait été mal estimée en début de projet. Cette estimation initiale a été faussée par un optimisme exagéré basé sur l'argument commercial d'un développement graphique simple de type « Drag and Drop ». Dans les faits, il s'est avéré que, comme pour tout nouveau langage, l'apprentissage est long et fastidieux. De plus, il convient d'appréhender les commandes Powershell pour une conception optimale des jobs séquence.

Cette phase m'a également permis de faire évoluer les documents produits lors de la phase précédente (matrice de chiffrage, modèle de spécification et guide des bonnes pratiques). Ces documents avaient été, dans un premier temps, élaborés sur des connaissances théoriques. J'ai pu, durant cette phase de réalisation, les confronter à la pratique.

Enfin, cette phase m'a permis d'évaluer concrètement les possibilités offertes par Datastage et de lister les évolutions et projets que nous pourrions menés dans le cadre de la rationalisation du SI.

VI Le pilotage du projet

L'organisation et la conduite de ce projet suit la procédure de gestion de projet existante chez CGA. J'ai fait également le choix de suivre certaines préconisations du PMBOK afin d'améliorer cette conduite de projet qui peut parfois être un peu simpliste dans certains aspects.

VI.1 Le lancement du projet

Le lancement est l'étape initiale d'un projet. Elle permet d'identifier les acteurs du projet, de définir son périmètre et de choisir le niveau de suivi à appliquer pendant la vie du projet. Chez CGA, cette phase est formalisée par une réunion de lancement à laquelle doivent participer l'ensemble des parties prenantes identifiées. A l'issue de cette réunion, la fiche de lancement du projet, le planning prévisionnel et la matrice des responsabilités sont rédigés et transmis à l'ensemble des acteurs.

Un des éléments majeurs dans la gestion du projet est l'identification nominative des différents acteurs et de leurs rôles. Dans le cadre du projet présent, les acteurs suivants sont identifiés :

- **le délégataire** : il est le commanditaire du projet. Il lui appartient d'exprimer son besoin et de justifier l'intérêt (économique et/ou stratégique) du projet. Il est responsable de l'acceptation finale des travaux. Il prend les décisions de fond concernant le projet (sur proposition des autres acteurs) et rend les arbitrages nécessaires. Le délégataire de ce projet est Mme Baras (DSI de CGA).
- **le responsable métier** : choisi par le délégataire, il est la référence fonctionnelle sur le sujet étudié. Il lui appartient de décrire et de définir précisément les besoins à couvrir. Il est garant de la cohérence du périmètre étudié, et il s'appuie à chaque fois que nécessaire sur des experts. Il participe à la validation des documents et signe le PV de recette. Les responsables métier de ce projet sont Mme Bataille (Responsable du service « Etudes ») et M. Lavedrine (son adjoint).

- **le chef de projet (MOA)** : il est responsable de la coordination de l'ensemble des intervenants et des activités. Il élabore et suit le planning du projet, anticipe les points de blocage éventuels et effectue les relances nécessaires au bon déroulement de l'ensemble. Il constitue et met à jour le dossier du projet sur une arborescence commune. Il rédige le cahier des charges à destination de la MOE. Il est responsable du bon déroulement de la recette. Ce projet étant considéré comme un projet DSI, je tiens ce rôle dans le cadre de ce projet.
- **le maitre d'œuvre (MOE)** : il est responsable de la réponse aux besoins exprimés par le maitre d'ouvrage. Il lui appartient de proposer les moyens appropriés pour obtenir les résultats souhaités, d'établir les devis correspondants, et de mener à bien les travaux pour livrer la solution, définie conjointement, dans le respect des délais et du budget. Il est responsable des moyens et de l'organisation à mettre en place au sein de la DSI pour atteindre le but fixé. Il est responsable de la production du livrable final du projet. Je tiens également ce rôle dans le cadre ce projet.
- **le chef de projet infrastructure** : il est nommé en début de projet. Il définit, selon les standards techniques et sécurité, la solution d'architecture technique en collaboration avec la MOE pour les projets fonctionnels. Il participe et valide le dossier d'exploitation en collaboration avec le chef de projet MOE. Il est responsable de fournir l'infrastructure nécessaire au bon déroulement du projet (dont l'environnement de recette). Le chef de projet infrastructure de ce projet est M. Tlili.
- **le correspondant sécurité des systèmes d'information (CSSI)** : il intervient dès le début du projet. Il émet ses préconisations et in fine, valide la conformité de la solution du point de vue sécurité du SI. Il s'assure que le projet répond aux exigences de sécurité définit dans le PGSSI (Politique Globale de Sécurité des Systèmes d'Information). Il valide les documents techniques du projet. Il participe à l'élaboration du dossier de sécurité et valide avec les instances définies. Il recourt autant que de besoin au RSSI (Responsable Sécurité des Systèmes d'Information) pour validation et arbitrage. Le CSSI de ce projet est M. Maudoux.
- **les analystes et programmeurs** : souvent regroupés les deux types de ressources répondent à l'étude ou l'analyse détaillée du projet et à sa réalisation dans le respect des plannings fournis par le chef de projet MOE. Dans la plupart des travaux de ce

projet, je tiens ces rôles mais j'intègre également les analystes et programmeurs de la DSI pour les sensibiliser au produit.

- **les prestataires et partenaires** : ces ressources externes à CGA, peuvent intervenir dans la mise en œuvre ou la réalisation de tout ou partie de la solution. Elles interviennent sous le contrôle du chef de projet MOE/MOA dans le respect des plannings établis avec ce dernier. Dans ce projet, la société NSIT intervient en tant qu'intégrateur de la solution. Cette société n'est pas citée lors de la phase de lancement car le choix de l'intégrateur fait partie intégrante du projet.

La fiche de lancement, signée par le délégataire, liste les différents acteurs du projet. Elle permet aussi de préciser la description du projet et sa justification. Elle détermine également le suivi du projet. Il appartient au responsable du projet de choisir la « granularité » du suivi qui semble appropriée à un contexte donné. Dans le cadre de mon projet, j'ai choisi d'animer un comité de projet mensuel avec l'ensemble des acteurs identifiés. Enfin, un lien vers la matrice des responsabilités est également intégré à cette fiche.

La matrice des responsabilités (cf. Annexe 5) permet de définir pour chaque projet la liste des documents qui seront à produire et les responsabilités de chaque acteur vis-à-vis de ces livrables.

La phase de lancement détermine également à un macro-planning prévisionnel qui pourra faire l'objet de révisions en cours de projet.

<i>Phases du projet</i>	<i>Jalons prévisionnels</i>
Lancement du projet	01/12/13
Choix de la solution	31/01/14
Mise en place de la solution	30/04/14
Paramétrages Lot 1	31/05/14
Paramétrage Lot 2	10/08/14
Clôture du projet	31/08/14

Tableau 10 : Macro-Planning prévisionnel du projet

VI.2 Le suivi du projet

Tout au long du projet et de ses différentes phases, j'ai utilisé des outils me permettant de suivre au plus près les performances du projet et ses éventuelles dérives.

Quotidiennement, je maintiens un document de reste à faire (cf. Annexe 10). Ce document consiste à suivre les charges restantes pour chaque tâche du projet. Chacune des tâches est découpée en 3 phases : Analyse/Spécification, Réalisation/Tests unitaires, Recette Technique. Je mets à jour pour chaque tâche un pourcentage d'avancement qui s'applique sur la charge initialement chiffrée et calcule donc un reste à faire. Je renseigne également pour chaque tâche une date de début, une date de fin et un acteur responsable de la tâche.

Hebdomadairement, je mets à jour une fiche de suivi de projet (cf. Annexe 9). Cette fiche est mise à disposition de l'ensemble des acteurs du projet et fait donc office de moyen de communication sur l'avancement du projet. Elle contient les sections suivantes :

- **l'état général** : il s'exprime sous forme d'un smiley permettant de donner le sentiment actuel de manière subjective sur le déroulement du projet.
- **l'avancement du projet** : cette section liste les principales tâches réalisées ou en cours de réalisation sur le projet.
- **les risques / problèmes à traiter** : cette section décrit les risques rencontrés jusqu'à présent dans le projet.
- **les décisions prises / attendues / à prendre** : cette section stipule les décisions prises par le chef de projet ou les décisions à prendre au niveau supérieur. Ces décisions attendues seront soumises au délégataire ou aux responsables concernés lors des comités de projet.
- **les principaux jalons** : la liste des jalons présente les dates initialement planifiées, les dates potentiellement révisées et les dates de réalisation effectives.
- **le suivi des charges** : les charges sont présentées par lot avec pour chacun une mise en perspective entre la charge initiale, le réalisé et le reste à faire. Ce suivi permet de calculer le pourcentage d'avancement et également le pourcentage de dérive.

Mensuellement, j'organise un comité de projet réunissant l'ensemble des acteurs du projet. Le comité de projet dure une heure et suit toujours le même déroulé :

- un rappel des objectifs de la période écoulée
- un focus sur une ou plusieurs des réalisations de la période écoulée
- la présentation de la dernière fiche de suivi de projet
- la présentation des décisions à prendre pour arbitrage
- la liste des objectifs de la période à venir

Chacun comité de projet donne lieu à un compte rendu diffusé aux acteurs du projet et plus largement mis à disposition de l'ensemble de la DSI.

Pendant toute la durée du projet, j'ai porté une attention particulière à la conduite du changement. Pour cela, j'ai essayé d'impliquer au maximum les futurs utilisateurs de la solution notamment en leur proposant des présentations ciblées de l'outil et en luttant contre les résistances et les fantasmes relatifs à un nouvel outil. La délégataire du projet m'a également bien aidé dans l'implication des équipes en intégrant dans les objectifs annuels de chaque chef de projet la mise en place un flux ETL.

VI.3 La clôture du projet

La clôture est l'étape finale d'un projet. Elle fait l'objet d'un comité de clôture regroupant l'ensemble des acteurs du projet (internes CGA et externes). L'objectif de cette réunion est de faire le bilan du projet et de constater la fin des travaux associés.

A l'issue de cette réunion, la fiche de fin de projet est remplie et diffusée aux différents acteurs. Cette fiche contient uniquement un bilan résumé, la date de début d'amortissement comptable et la signature du délégataire et du maître d'œuvre.

La clôture du projet doit aussi faire l'objet d'un document de bilan. Cette étape est systématiquement ignorée chez CGA en raison de la surcharge ambiante qui implique de passer au projet suivant le plus rapidement possible. J'ai donc mis un point d'honneur à réaliser ce document qui pourra également servir de modèle pour les projets suivants. J'ai organisé ce document selon les chapitres suivants :

- **la performance du projet** : ce chapitre liste notamment les accomplissements majeurs et les charges consommées. Il décrit également la performance en termes de points positifs et de points négatifs (cf. chapitre suivant : le bilan du projet).
- **les retours d'expérience phase par phase**
- **les bonnes pratiques à capitaliser** : ce chapitre est primordial dans le cadre d'une amélioration continue des projets chez CGA. J'ai notamment identifié trois bonnes pratiques de ce projet qui sont à reconduire dans les futurs projets.
 - nommer un chef de projet infrastructure sur tous les projets d'acquisition de solution. Cela permet au chef de projet MOE d'avoir un interlocuteur dédié pour les parties Infrastructure/Production. Ce n'est pas toujours le cas actuellement dans les projets SI et cela provoque parfois des tensions et des problèmes de communication.
 - élaborer un business case permettant de mettre en relief les gains attendus par la mise en place du projet. Ce document que j'ai élaboré n'est pas une habitude chez CGA. Il s'agit pourtant des arguments les plus importants pour la direction générale.
 - élaborer un manuel utilisateur (guide des bonnes pratiques) permettant une uniformité dans l'utilisation de l'outil par l'ensemble des intervenants.

Aujourd'hui, trop d'outils sont utilisés chez CGA avec des pratiques différentes selon les intervenants. Le résultat de ces utilisations personnalisées est une maintenance compliquée.

- **les opportunités d'amélioration** : ce chapitre décrit les pratiques qui n'ont pas été suivies dans ce projet mais qui aurait pu être bénéfiques et sont recommandées pour les futurs projets. J'ai principalement constatés deux pistes principales :
 - réaliser un POC dans le cadre d'un nouveau produit. Je recommande de réaliser une maquette avant d'acquérir un nouvel outil. Cela permet, sur un périmètre restreint, de constater objectivement les fonctionnalités d'un outil et mettre en perspective la réalité par rapport au discours commercial.
 - déployer 3 environnements distincts. La mise en place d'un environnement de développement, d'un environnement d'homologation et d'un environnement de production fiabilise les livraisons et les recettes. Dans le cadre de ce projet, nous avons choisi de disposer de 2 environnements seulement (non-production et production) pour des raisons budgétaires (coûts des licences).

Enfin, la clôture du projet implique également la fermeture de la ligne budgétaire associée au projet. Cela signifie qu'aucune charge ne pourra plus être consommée sur ce projet.

VII Le bilan

VII.1 Le bilan du projet

Le bilan du projet est globalement positif. Les points positifs que l'on peut ressortir sont les suivants :



l'outil Datastage est installé. L'objectif principal du projet de doter CGA d'un outil ETL pour l'intégration des données est atteint. Datastage est fonctionnel dans le SI de CGA avec un environnement de production et un environnement de non-production.



le délai du projet a été globalement respecté. La fin de projet était initialement prévue le 31 août 2014. En effet, à partir du 1^{er} septembre, je devais prendre mes nouvelles fonctions. Dans les faits, j'ai pu finaliser le projet courant septembre (recette des derniers développements du lot 2, comité de clôture et rédaction du bilan de projet).



les contraintes d'intégration dans le SI sont respectées. L'outil est totalement intégré au SI grâce à la prise en compte des contraintes globales : gestion des droits par l'intermédiaire de l'Active Directory, supervision grâce à des requêtes exécutables par notre outil de monitoring centralisé, exécution des jobs par notre scheduler grâce au script développé, automatisation de la livraison par notre outil de déploiement grâce au script développé.



les principaux documents ont été livrés : Les livrables identifiés lors du lancement du projet ont été rédigés (manuel utilisateur, dossier d'architecture...). Certains de ces livrables (notamment le document des bonnes pratiques) seront amenés à évoluer avec la montée en compétence des équipes sur l'outil.



des développements sont livrés ou en cours : Les développements du lot 1 et du lot 2 sont livrés (à l'exception du remplacement de SETI-Map toujours en stand-by). Plusieurs autres projets en cours font déjà l'objet de flux ETL ce qui démontre bien l'acceptation du produit.

En revanche, comme dans tout projet, on peut identifier des points négatifs :



peu de développements prévus dans le lot 2 ont été réalisés : Le nombre de développements inscrits dans le lot 2 en début de projet s'est avéré trop ambitieux. Le délai étant incompressible, nous avons donc joué sur ce levier en supprimant certains des développements qui consistaient à réécrire des flux existants dans une optique de rationalisation. Ces flux seront intégrés dans le projet de rationalisation budgété pour l'année 2015.



l'apprentissage de l'outil est plus complexe qu'escompté : Les charges initialement prévues en termes de développement ont dû être révisées en cours de projet car elles étaient basées sur des prévisions trop optimistes. Il s'avère que les développements de flux Datastage nécessitent une bonne connaissance de l'outil. La prise en main n'est pas plus simple que celle nécessaire à l'apprentissage d'un nouveau langage.



le script d'exécution des flux n'est pas totalement satisfaisant : Le script mis en place actuellement ne renvoie qu'un statut OK ou KO. Un travail supplémentaire doit être réalisé pour affiner ce code retour en fonction des exceptions et des erreurs. Cette tâche sera affectée à une ressource du service Infrastructure et Production travaillant sur le projet de mise en place du scheduler.



les connaissances en interne sont encore trop légères : Les formations complètes ont été actuellement suivies uniquement par les chefs de projet interne. Les développeurs qui utilisent Datastage quotidiennement ont seulement suivi une journée de passage de connaissances que j'ai organisé avec l'intégrateur de la solution. Pour remédier à ce problème, des formations supplémentaires seront planifiées en 2015. De plus, deux ressources expertes sur Datastage (un chef de projet et un développeur) ont été intégrées au service « Etudes » suite à ce projet.

En termes de budget, le bilan est le suivant :

	<i>Prévisionnel</i>	<i>Consommé</i>
Licence	51 000	51 000
Formation	15 600	40 400
Services	27 885	55 835
J/H CGA		
Etudes	240	160
Formation	12	28
I&P	6	14

Tableau 11 : Bilan budgétaire du projet

On constate inévitablement des différences entre le prévisionnel et le consommé mais celles-ci peuvent s'expliquer :

- **formation** : nous avons prévu, initialement, de former 3 personnes du service Etudes. Nous avons révisé notre jugement en cours de projet et avons décidé de former 5 personnes du service Etudes et 2 personnes du service Infrastructure & Production. De plus, nous avons organisé avec notre intégrateur une journée de sensibilisation en interne pour les développeurs du service Etudes.
- **services** : les coûts de service comprenaient, dans la proposition initiale, une prestation d'expertise et installation et une prestation d'accompagnement au développement. Nous avons choisi de compléter la seconde prestation de 10 jours et de poursuivre la montée en compétence des équipes avec l'intervention d'un chef de projet expert pendant 30 jours.
- **J/H CGA** : la charge en J/H « Etudes » a été moins conséquente qu'escomptée. En effet, le périmètre du lot 2 a évolué en cours de projet car certains paramétrages ont été affectés directement aux projets les concernant. En revanche, la charge « I&P » a été largement plus lourde que prévue en raison principalement des nombreuses difficultés rencontrées lors de la phase d'installation.

VII.2 Le bilan personnel

D'un point de vue personnel, j'estime que ce projet est un succès puisqu'il a globalement répondu aux objectifs fixés lors de son lancement.

Pour la première fois, en huit années de présence chez CGA, on m'a permis de travailler, pendant 9 mois, exclusivement sur un projet. Cette opportunité est très rare dans le contexte de CGA qui est extrêmement dynamique et qui impose aux chefs de projet de travailler sur de nombreux projets en parallèle. Cette exclusivité m'a permis de travailler en respectant une méthodologie classique mais qui est souvent réduite à son expression minimum dans les projets CGA en raison de la surcharge. J'ai constaté que le respect des différentes phases d'un projet et le fait de pouvoir se concentrer sur un sujet unique sont bien entendu très bénéfiques au bon déroulement de celui-ci.

J'ai également profité de ce projet pour travailler sur les points faibles que je me connais et que l'on m'a déjà signalé par le passé. En particulier, j'ai mis l'accent sur la communication qui me fait parfois défaut et sur l'affirmation de mes convictions. J'ai notamment du faire preuve de fermeté vis-à-vis des fournisseurs pour éviter certaines dérives ainsi qu'avec mes collègues pour éviter de me faire phagocytter par des tâches extérieures à mon projet. Mon succès a été moins probant sur un autre défaut récurrent sur lequel je dois continuer à travailler : mon optimisme. Ce trait de caractère m'amène parfois, et notamment dans mes chiffrages, à ne pas prendre en compte tous les risques et les contraintes d'un projet et de son environnement.

Ma conclusion personnelle de ce projet est également que les difficultés et les challenges sont souvent dans « l'humain » et plus rarement dans la technique. Fort de ce constat, je prends en charge mon nouveau poste avec un rôle de management d'équipe.

La clôture de ce projet après un cursus de 12 ans au CNAM me provoque un double sentiment. D'un côté, j'éprouve un soulagement d'avoir été jusqu'au bout d'un parcours parfois éprouvant. De l'autre, je redoute un sentiment de « manque » après toutes ces années pendant lesquelles mon esprit était toujours plus ou moins préoccupé par le CNAM.

VII.3 Les perspectives

Grâce à ce projet, CGA s'est doté d'un outil ETL performant et réputé. Cet investissement, couplé aux nombreux projets de rationalisation en cours, ouvre de nouvelles perspectives au SI de CGA.

En parallèle à la clôture du projet, j'ai listé et macro-chiffré les flux existants susceptibles d'être intégrés dans l'ETL pour simplifier et optimiser notre SI. Cet inventaire de flux sera pris en compte dans la Road Map 2015 et affecté à la ligne budgétaire « Rationalisation / Simplification du SI ». Ces tâches seront affectées à un nouveau pôle « Rationalisation et Outils Transverses » créé dans le cadre de la réorganisation du service « Etudes ».

J'ai également listé les fonctionnalités de Datastage non encore utilisées chez CGA mais qu'il conviendra de garder à l'esprit pour les futurs développements et la modernisation du SI. Parmi celles-ci, j'ai mis en perspectives les possibilités offertes pour le traitement des flux XML actuellement peu utilisés chez CGA mais qui vont le devenir en raison notamment de l'évolution des formats bancaires internationaux du format plat vers le format XML (dans le cadre du SEPA notamment).

Parmi les nombreux projets majeurs composant la Road Map 2015 de CGA, il conviendra de toujours privilégier l'utilisation de Datastage, qui devient d'ors et déjà une brique centrale du SI, lors de la détermination de l'architecture. Ainsi, le projet de refonte du site transactionnel communiquera avec la plateforme centrale via des jobs ETL. Le projet de mise en place d'un infocentre consiste à créer un véritable Datawarehouse destiné au décideur. Ce Datawarehouse sera alimenté par des jobs Datastage et la restitution sera effectuée par Cognos qui est également un outil IBM. Un projet de renouvellement de notre CRM est également prévu. Ce progiciel (probablement Salesforce) sera la source ou la cible de nombreux flux ETL pour s'intégrer parfaitement dans le SI de CGA. Enfin, la mise en place d'un référentiel réglementaire permettant de regrouper les indicateurs de nos différentes filiales est également au programme pour 2015. Ce référentiel sera alimenté par un grand nombre des applications connexes existantes et sera donc, de fait, la cible de nombreux jobs Datastage.

Tous ces projets ambitieux demanderont certainement un renfort en termes de ressources. Le fait d'avoir choisi Datastage nous ouvre aussi des perspectives en tant que choix des profils susceptibles d'intégrer la DSI.

Conclusion

Le projet de mise en place d'une solution ETL au sein du SI de CGA est terminé. La durée imposée de 9 mois a été respectée. Les leviers « coût » et « fonctionnalités » ont évolués au cours du projet.

Comme dans tout projet, j'ai rencontré des difficultés mais celles-ci ont pu être surmontées grâce aux efforts menés par l'ensemble des acteurs impliqués. Ce projet m'a permis de compléter mon expérience notamment en termes de communication et de méthodologie. Il m'a également permis d'acquérir de nouvelles compétences dans le domaine de l'acquisition d'une solution (rédaction d'un appel d'offre, choix d'une solution basé sur des critères objectifs, suivi d'un fournisseur...).

La solution ETL étant récente chez CGA, du travail sera encore nécessaire pour que celle-ci soit complètement intégrée au SI mais je suis plutôt satisfait de l'accueil qui est fait à cet outil par les utilisateurs de la DSI. Malgré quelques réticences observées et inhérentes à la peur du changement, les développeurs et les chefs de projet sont plutôt moteurs sur le sujet.

Ce projet me sera extrêmement bénéfique dans la conduite des projets qui me seront confiés en 2015 puisque ceux-ci intégreront nécessairement des briques ETL dans leur architecture technique.

Je suis également satisfait du choix de CGA de s'orienter vers une rationalisation et une modernisation du SI qui, j'en suis convaincu, apporteront des bénéfices conséquents à l'entreprise et notamment à la DSI qui souffre régulièrement de la complexité et de l'obsolescence du SI.

Références

Bibliographie

- [01] **Guide du corpus des connaissances en management de projet (Guide PMBOK)**. 4^{ème} édition. Newtown Square : Project Management Institute, 2008, 459 p.
- [02] J.P. Deschanel et Laurent Lemoine : **Que Sais-Je ? L'affacturage**. 2^{ème} édition. Presses Universitaires de France, 1993, 130 p.
- [03] Nagraj Alur, Celso Takahashi, Sachiko Toratani, Denis Vasconcelos : **IBM Infosphere Datastage Data Flow and Job Design**. IBM Redbooks Publication, 2008, 658 p.
- [04] Julius Lerm, Paul Christensen : **Infosphere DataStage Parallel Framework Standard Practices**. IBM Redbooks Publication, 2010, 458 p.
- [05] Benoit Kaluzni : **Intégration de données financières avec un outil ETL Open Source**. Mémoire CNAM, 2014, 117 p.

Webographie

- [06] DSXchange : <http://www.dsxchange.com>
- [07] Guide comparatif ETL : <http://www.guidescomparatifs.com/etl-integration.asp>
- [08] Documentation IBM :
<http://www-01.ibm.com/support/docview.wss?uid=swg27035772>
- [09] Wikipedia : http://fr.wikipedia.org/wiki/Extract_Transform_Load
- [10] Talend : <https://fr.talend.com/>

Normes

- [12] NF EN ISO 9001 – Systèmes de management de la qualité. Afnor, 2008, 28 p.
- [13] ISO 14001:2004 – Systèmes de management environnemental. Afnor, 2004, 25 p.

Table des annexes

Annexe 1 Adéquation aux besoins	106
Annexe 2 Analyse des coûts.....	108
Annexe 3 Classement des réponses.....	109
Annexe 4 Business Case	111
Annexe 5 Format du fichier Monoprix en entrée	113
Annexe 6 Format du fichier Monoprix en sortie.....	114
Annexe 7 Format du fichier d'items multi-format	115
Annexe 8 Matrice des responsabilités.....	116
Annexe 9 Fiche de Suivi de Projet	118
Annexe 10 Reste A Faire	119

Annexe 1

Adéquation aux besoins

Id Besoin	Libellé Besoin	Fournisseur X	IBM	Fournisseur Y
Besoins vitaux				
Plateforme				
1.1	Plateforme sous Windows	Oui	Oui	Oui
1.2	BD référentiel sous SQL Server	Oui	Oui	Oui
1.3	Intégration dans la solution de supervision existante (TANGO04)	Oui	Oui	Oui
1.4	Intégration dans le scheduler existant (ROBOT) ou futur	Oui	Oui	Oui
1.5	Intégration dans la solution de versioning (packages)	Oui	Oui	Oui
Connecteurs				
2.1	Lecture / Ecriture SQL Server (connecteur natif)	Non	Oui	Oui
2.2	Lecture / Ecriture DB2 AS400 (connecteur natif)	Non	Oui	Oui
2.3	Lecture / Ecriture fichiers textes (plat, CSV)	Oui	Oui	Oui
2.4	Lecture / Ecriture fichiers XML	Oui	Oui	Oui
2.5	Connexion par connecteurs génériques (ODBC / JDBC)	Oui	Oui	Oui
Traitements spécifiques				
3.1	Scan d'un répertoire (Windows ou IFS)	Oui	Oui	Oui
3.2	Scan d'une bibliothèque AS400	Non	Oui	Non
3.3	Dézippage / Zippage	Oui	Oui	Oui
3.4	Lecture / Envoi de mail	Oui	Oui	Oui
Exploitabilité / Administration				
4.1	Transfert via FTP, sFTP	Non	Oui	Oui
4.2	Parallélisation des traitements	Oui	Oui	Oui
4.3	Connexion via LDAP	Non	Oui	Oui
4.4	Support d'une charge quotidienne de 15 à 20 Go	Oui	Oui	Oui
4.5	Distinction Paramétrage (Développement) et Tuning (Exploitation)	Oui	Oui	Oui
4.6	Système de Sauvegarde – Reprise	Oui	Oui	Oui
Besoins secondaires				
0.01	Lancement d'un CALL AS400	Non	Non	Non
0.02	Lecture fichier Excel	Non	Oui	Oui

0.03	Lecture fichier PDF	Non	Non	Oui
0.04	Connexion SSO	Non	Non	Oui
0.05	Transfert via d'autres protocoles (CFT...)	Non	Non	Non
0.06	Bibliothèque de standards industriels (SEPA, SWIFT...)	Oui	Oui	Oui
0.07	Connecteur COGNOS (futur évolution du Datawarehouse)	Non	Oui	Non
0.08	Analyse d'impact		Oui	Oui
0.09	Rapport d'intégration		Oui	
0.10	Génération automatique de documentation		Oui	

Annexe 2

Analyse des coûts

	Fournisseur X	IBM	Fournisseur Y
Licences			
<i>Environnement de production</i>	48 000 €	26 642 €	113 330 €
<i>Environnement de développement/ test</i>	24 000 €	17 520 €	29 580 €
<i>Environnement de backup</i>	0 €	0 €	0 €
<i>Options retenues</i>	0 €	6 832 €	30 560 €
TOTAL PRODUIT	72 000 €	50 994 €	173 470 €
Prestation d'intégration	24 000 €	27 885 €	31 000 €
	(20J)	(13J + 26J)	(25J)
Formation (sur une base de 3 personnes)	12 000 €	15 000 €	10 000 €
TOTAL SERVICE	36 000 €	42 885 €	41 000 €
Maintenance (annuel)	15 840 €	11 500 €	35 000 €
		(1ère année incluse dans la licence)	
TOTAL MAINTENANCE (3 ANS)	47 520 €	23 000 €	105 000 €

Annexe 3 Classement des réponses

ETL - CLASSEMENT DES CONSULTES

	Coef / Poids	Fournisseur X	IBM	Fournisseur Y
Respect des procédures	1	10,00	12,00	9,00
Délai de réponses		4,0	3,0	2,0
Forme de la réponse		3,0	4,0	2,0
Synthèse de la proposition		1,0	1,0	1,0
Références (contacts)		2,0	4,0	4,0
Contenu de la proposition	2	8,00	12,00	10,00
Présentation / soutenance		3,0	2,0	4,0
Méthodologie proposée		1,0	4,0	1,0
Critères financiers	3	12,00	18,00	6,00
Coûts du projet		2,0	3,0	1,0
Coûts récurrents		2,0	3,0	1,0
Analyse du produit	5	205,00	265,00	275,00
Plateforme		4,0	4,0	4,0
Fichiers supportés		2,0	4,0	4,0
Connecteurs d'accès aux données		2,0	4,0	4,0
Prise en main de l'outil (complexité de mise en œuvre)		3,0	3,0	3,0
Transformations disponibles		2,0	4,0	4,0
Déclenchement des jobs		3,0	4,0	4,0
Traitements spécifiques		2,0	4,0	4,0
Gestion des configurations		4,0	4,0	4,0
Architecture		4,0	4,0	4,0
Gestion des droits d'accès (SSO)		4,0	3,0	4,0

Adéquation avec la matrice des besoins		3,0	4,0	4,0
Rationalisation SI (applications supprimées)		3,0	3,0	4,0
Performances		2,0	4,0	4,0
Exploitabilité		3,0	4,0	4,0
Services proposés	4	36,00	36,00	36,00
Support		3,0	3,0	3,0
Formation		3,0	3,0	3,0
Assistance à la mise en place		3,0	3,0	3,0
Soutenance / Suivi	2	4,00	8,00	6,00
Soutenance		2,0	4,0	3,0
TOTAL		275	351	342

Notation : 1 = très insuffisant, 2 = insuffisant, 3 = bon, 4 = très bon

Annexe 4 Business Case

CGA

Business Case - Projet ETL											
	2013	2014		2015		2016		2017		2018	
		Sans Datastage	Avec Datastage	Sans Datastage	Avec Datastage	Sans Datastage	Avec Datastage	Sans Datastage	Avec Datastage	Sans Datastage	Avec Datastage
Coûts du projet											
Licences	0	0	50994	0	0	0	0	0	0	0	0
Prestation	0	0	48685	0	0	0	0	0	0	0	0
Infrastructure	0	0	0	0	0	0	0	0	0	0	0
Mise en place	0	0	20000	0	0	0	0	0	0	0	0
Maintenance Logicielle	0	0	0	0	11500	0	11500	0	11500	0	11500
Développements											
Démarrage commerciaux	0	49500	16500	49500	16500	49500	16500	49500	16500	49500	16500
Outil de test de fichiers	0	0	27500	0	0	0	0	0	0	0	0
Migration new CGA-Contact	0	8250	2750	0	0	0	0	0	0	0	0
Chargement new CGA-Contact	0	16500	5500	0	0	0	0	0	0	0	0
Allemagne (GL)	0	4950	1650	0	0	0	0	0	0	0	0
Simplification / Rationalisation											
Remplacement SETI-Map	0	0	11000	0	0	0	0	0	0	0	0
Datawarehouse	0	0	16500	0	0	0	0	0	0	0	0
Flux opérations spéciales	0	0	16500	0	0	0	0	0	0	0	0
APF	0	0	27500	0	0	0	0	0	0	0	0
Syndication	0	0	5500	0	0	0	0	0	0	0	0
Archivage Comptable	0	0	11000	0	0	0	0	0	0	0	0
PCA Financement	0	0	11000	0	0	0	0	0	0	0	0
Blanchiment	0	0	11000	0	0	0	0	0	0	0	0
INSEE	0	0	5500	0	0	0	0	0	0	0	0
Ligis	0	0	5500	0	0	0	0	0	0	0	0
Relance	0	0	5500	0	0	0	0	0	0	0	0
Rapprochement Bancaire	0	0	27500	0	11000	0	0	0	0	0	0
Maintenance applicative											
Chargement New CGA-Contact	0	0	0	1650	544,5	1650	544,5	1650	544,5	1650	544,5
Outil de test de fichiers	0	0	0	0	907,5	0	907,5	0	907,5	0	907,5
Datawarehouse	11000	11000	3630	11000	3630	11000	3630	11000	3630	11000	3630
Flux opérations spéciales	16500	16500	16500	16500	5445	16500	5445	16500	5445	16500	5445
APF	11000	11000	11000	11000	3630	11000	3630	11000	3630	11000	3630
Syndication	12650	12650	12650	12650	4174,5	12650	4174,5	12650	4174,5	12650	4174,5
Archivage Comptable	1100	1100	1100	1100	363	1100	363	1100	363	1100	363
PCA Financement	1100	1100	1100	1100	363	1100	363	1100	363	1100	363
Blanchiment	1100	1100	1100	1100	363	1100	363	1100	363	1100	363
INSEE	550	550	550	550	181,5	550	181,5	550	181,5	550	181,5
Ligis	550	550	550	550	181,5	550	181,5	550	181,5	550	181,5
Relance	550	550	550	550	181,5	550	181,5	550	181,5	550	181,5
Rapprochement Bancaire	2750	2750	2750	2750	907,5	2750	907,5	2750	907,5	2750	907,5
Maintenance logicielle											
SETI-Map	12000	12000	12000	12000	0	12000	0	12000	0	12000	0
Exploitation											
Test de fichiers	50000	50000	0	50000	0	50000	0	50000	0	50000	0
TOTAL	120850	200050	391059	172000	59872,5	172000	48872,5	172000	48872,5	172000	48872,5

TOTAL	Sans Datastage	Avec Datastage	Cumul Sans Datastage	Cumul Avec Datastage	ROI total
2014	200 050	391 059	200 050	391 059	-191 009
2015	172 000	59 873	372 050	450 932	-78 882
2016	172 000	48 873	544 050	499 804	44 246
2017	172 000	48 873	716 050	548 677	167 374
2018	172 000	48 873	888 050	597 549	290 501

Hypothèses retenues dans les calculs

Divers	
Maintenance pour un projet (en % de la charge initiale) / an	10
Gain en maintenance Développement VS ETL (en %)	33
Coût CDP ETL (100 J/H)	20000
TJM moyen d'un développeur	550

Démarrages commerciaux	
Nombre de démarrages commerciaux / an	10
Mise en place d'un nouveau flux (conversion + contrôles) sans ETL (en J/H)	9
Mise en place d'un nouveau flux (conversion + contrôles) avec ETL (en J/H)	3

Outil Test de fichiers	
Mise en place d'un outil de test de fichiers (en J/H)	50
1 ETP (test de fichiers) / an	50000

New CGA-Contact	
Mise en place des outils de migration sans ETL (en J/H)	15
Mise en place des outils de migration avec ETL (en J/H)	5
Mise en place du chargement quotidien des données sans ETL (en J/H)	30
Mise en place du chargement quotidien des données avec ETL (en J/H)	10

Allemagne (GL)	
Mise en place du transfert-intégration du GL sans ETL (en J/H)	9
Mise en place du transfert-intégration du GL avec ETL (en J/H)	3

SETI-Map	
Mise en place du remplacement SETI-Map (en J/H)	20
Coût maintenance annuelle SETI-Map	12000

Simplification / Rationalisation	Charge initiale du projet (en J/H)	Mise en place de la simplification (en J/H)
Datawarehouse	200	30
Flux Opérations Spéciales	300	30
APF	200	50
Syndication	230	10
Archivage Comptable	20	5
PCA Financement	20	3
Blanchiment	20	5
INSEE	10	3
Ligis (Extraction / Chargement)	10	3
Relance (Extraction / Chargement)	10	3
Rapprochement Bancaire	50	20

Annexe 5

Format du fichier Monoprix en entrée

Ordre	Nom	Libellé	Type	Commentaire
1	CodeDO	Code REVERSE	A(3)	A définir avec le DO, lors de la signature du contrat
2	PositionDec	Position de la décimale	N(1)	Par défaut, indiquez « 2 ». Nombre de chiffres après la virgule
3	Version	Version	A(2)	Indiquez « 01 »
4	FormatDate	Format des dates	A(10)	AAAAMMJJ ou AAAA/MM/JJ ou JJMMAAAA ou JJ/MM/AAAA. L'année est sur 4 positions, le mois et le jour sur 2
5	ReferenceDO	Nom ou Référence du DO	A(30)	A définir avec le DO, lors de la signature du contrat
6	NumBAP	N° du BAP	N(15)	N° du BAP défini par le DO
7	IdFournisseur	Identifiant fournisseur	A(20)	Identifiant fournisseur chez le DO
8	SIRET	N° SIRET	N(14)	
9	DateEnvoi	Date d'envoi à CGA	A(10)	Format défini par champ 5
10	DateItem	Date de la pièce	A(10)	Format défini par champ 5
11	DateEcheance	Date d'échéance	A(10)	Format défini par champ 5 Date d'échéance est toujours renseignée, les avoirs sont affectés sur une échéance
12	RefItem	Numéro de la pièce	A(15)	
13	NumeroOrdre	Numéro d'ordre	N(5)	Par défaut « 00000 ». Utiliser dans le cas d'une facture multi-échéance. Cf. Annexe 2.
14	LibelleTypePiece	Libellé du type de pièce	A(30)	Exemple : Facture ou Avoir ou Note de débit.....
15	MontantTTC	Montant TTC	N(15)	Défini par la donnée 3 – Montant > 0
16	MontantEscompte	Montant escomptes TTC	N(15)	Défini par la donnée 3 – Montant >= 0
17	MontantAREgler	Montant à régler	N(15)	Défini par la donnée 3, Montant > 0 Montant à régler = Montant TTC – Montant escomptes TTC
18	Sens	Sens de la pièce	A(1)	« + » si 1 , « - » si -1
19	Devise	Code devise	A(3)	Devise norme ISO3

Annexe 6

Format du fichier Monoprix en sortie

Ordre	Nom	Libellé	Type	Règle/Alimentation
1	TypeFichier	Type de fichier	A(4)	Par défaut « ITEM »
2	Version	Version	A(2)	Par défaut « 01 »
3	CodeDO	Code SCF DO	A(3)	Code donné par le SAS SCF au DO
4	IdFournisseur	Identifiant fournisseur	A(30)	Référence du fournisseur chez le DO
5	TypeId	Type identifiant	A(3)	« SIR » pour SIRET « TVA » pour la TVA Intracommunautaire « RCS » autre
6	IdSociete	Identifiant société	A(50)	
7	Pays	Code pays ISO	A(2)	
8	NumBAP	Numéro de BAP	N(9)	
9	DateEnvoi	Date envoi du DO vers SAS	N(8)	Format AAAAMMJJ
10	DateItem	Date de l'item	N(8)	Format AAAAMMJJ
11	DateEcheance	Date d'échéance	N(8)	Format AAAAMMJJ
12	RefItemDO	Référence de l'item du DO	A(20)	
13	RefItemFournisseur	Référence de l'item du supplier	A(20)	
14	RefFactureLiee	Référence de la facture liée	A(30)	Référence de la facture dans le cas d'un avoir
15	Statut	Statut des items	N(3)	« 001 » Item VALIDE, « 002 » NON VALIDE
16	MontantTTC	Montant TTC de l'item (1)	N(17,4)	
17	MontantEscompteTTC	Montant de l'escompte (2)	N(17,4)	
18	MontantARegler	Montant à régler (3)	N(17,4)	(3) = (1) - (2)
19	Sens	Sens de l'item	A(1)	« + » pour les factures, « - » pour les avoirs
20	Devise	Devise ISO	A(3)	Code ISO
21	Libelle	Libellé de la pièce	A(200)	
22	IBAN	IBAN	A(50)	
23	BIC	BIC	A(15)	
24	Filler	Filler	A(500)	

Annexe 7

Format du fichier d'items multi-format

Format HEADER

Ordre	Libellé	Type	Règle/Alimentation
1	Segment Identifier	A(3)	« #H# »
2	Format	A(7)	« WEB_TAB »
3	Version	A(5)	« 00100 »
4	Message ID	A(32)	Référence du fournisseur chez le DO
5	Sending Client Id	A(10)	« SIR » pour SIRET « TVA » pour la TVA Intracommunautaire « RCS » autre
6	Sending Client Name	A(40)	
7	File Creation Date	D(8)	Format YYYYMMDD
8	Org Key Snd	A(20)	

Format PAYMENT

Ordre	Libellé	Type	Règle/Alimentation
1	Segment Identifier	A(3)	« #T# »
2	Request ID	A(16)	
3	Transaction Type	A(4)	« IPTC »
4	Currency	A(3)	
5	Sending Account Number	A(35)	
6	Receiving Client ID	A(10)	
7	Receiving Client Name	A(40)	
8	Receiving Account Number	A(35)	
9	Payment Due Date	D(8)	Format YYYYMMDD
10	Payment Amount	N(23)	Numérique avec 2 décimales
11	Org Key Rec	A(20)	

Format REFERENCE

Ordre	Libellé	Type	Règle/Alimentation
1	Segment Identifier	A(3)	« #R# »
2	Reference	A(65)	Ex : Belegnummer 323922 vom 07.10.2013 Amt:18731.71

Format FOOTER

Ordre	Libellé	Type	Règle/Alimentation
1	Segment Identifier	A(3)	« #F# »
2	Number of payment	A(8)	
3	Checksum Type	A(10)	
4	Checksum	A(32)	

Annexe 8

Matrice des responsabilités

Projet : ETL

V : Valide

P : Participe

E : Elabore

I : Pour information

Liste des documents	A produire sur le projet	Sponsor	MOA	MOE	CP Infrastructure	Utilisateur	CSSI
		MFBaras	YBataille MLavedrine	FCesbron	FTlili	FGhezali PLaudon XGuerin	XMaudoux
Expression de besoin (Cahier des charges)	Oui	V	E	V	I	P	V
Dossier de sécurité	Non						
Fiche de lancement	Oui	V	E	P	I	I	I
Chiffrage	Oui		I	E	P		
Matrice de décision	Oui	V	E	P	P	P	I
Spécification fonctionnelles générales	Oui	I	V	E	I	P	I
Spécifications fonctionnelles détaillées	Oui			E / V	P / V		I
Planning	Oui		E	E			
Fiche de suivi de Projet	Oui	I	E / V	E / V	P		P
Compte-rendu Comité de pilotage	Non						
Compte-rendu Comité opérationnel	Oui	V	E	P			

Document d'architecture technique	Oui			E	P / V		P / V
Dossier / Plan de tests	Oui		E	E		P	
Dossier d'exploitation	Oui			E	V		I
Manuel utilisateur	Oui		E			V	
Dossier de mise en production	Oui			E	V		
PV de recette	Oui		V	I	I	E	
Bilan de projet	Oui	V	E	P	P	P	P

Annexe 9

Fiche de Suivi de Projet

Projet : ETL
Date de mât : 22/07/2014
Nom du CP : F. Cesbron



Avancement du projet

- ✕ La solution retenue est la solution IBM (DataStage)
- ✕ L'installation de l'environnement de non-production est terminé
- ✕ Les paramétrages de la migration Monoprix sont terminés
- ✕ La recette de la migration Monoprix est terminée
- ✕ Le flux de la migration Monoprix est en production
- ✕ Les paramétrages du New Aquarius (Allemagne) sont terminés
- ✕ La recette des paramétrages du New Aquarius est terminée
- ✕ Le remplacement de SETI-Map est en cours de développement
- ✕ La rédaction des bonnes pratiques de développement est terminée
- ✕ La rédaction des bonnes pratiques d'exploitation est en cours de validation
- ✕ L'installation de la console operation est effective en DEV, reste à l'appliquer en PRD pour permettre la supervision via Tango04
- ✕ Formation développement suivie par FCesbron, LLeroyer et RLussiez
- ✕ Formation administration en cours pour FTilli et PCourdouan

Risques / Problèmes à traiter

L'apprentissage de l'outil est plus chronophage que prévu --> équivalent à l'apprentissage d'un nouveau langage

La rédaction des bonnes pratiques de développement et de production (gestion des erreurs, reprises, mise en production, supervision, déclenchement...) est primordiale et doit l'objet d'un consensus entre DSI/ETU et DSI/EXP --> ce consensus n'est toujours pas validé notamment concernant le périmètre de chaque outil (Bridge, Robot, Datastage...)

Etat général

Global	M-2	M-1	M

Décisions prises / attendues / à prendre

Choix de la solution pour les postes clients ?
La solution de client lourd installé par un script SCCM est prise, il reste à la mettre en oeuvre.

Choix dans la gestion des droits AD ?
Le choix entre la modification des noms de groupes (suppression des espaces) et la gestion par utilisateurs n'est toujours pas arrêtée

Principaux jalons (Dates)

Jalons	Planifiées	Révisées	Réalisées
Envoi de l'appel d'offre	15/11/13		15/11/13
Sélection (short-list)	19/12/13		19/12/13
Choix de la solution	31/01/14		13/02/14
Installation de la solution (DEV)	28/02/14	31/03/14	31/03/14
Installation de la solution (PRD)	30/04/14	31/05/14	25/06/14
Lot 1	31/05/14	15/06/14	30/06/14
Lot 2	10/08/14	31/08/14	

Type de ressource	Lot	Commentaire	Charge initiale	Charge ajoutée	Charge globale	Réalisé	Reste à faire	Révisé (Réalisé + Raf)	% Avancement	Dérive
DSI/ETU	Acquisition d'une solution		20		20	24	0	24	100%	-20%
DSI/ETU	Installation / Mise en place		10		10	21	0	21	100%	-110%
DSI/EXP	Installation / Mise en place		6		6	10	0	10	100%	-67%
DSI/ETU	Lot 1 : Paramétrage projets contraints		94	-68	26	22	4	26	85%	0%
DSI/ETU	Lot 2 : Paramétrage simplification briques existantes		86		86	10	76	86	12%	0%
DSI/ETU	Pilotage		20		20	10	10	20	50%	0%
Total :			236	-68	168	97	90	187	74%	-33%

Annexe 10

Reste A Faire

Statut : ECA = En cours d'analyse / ATE = Analyse terminée / ECD = En cours de développement / ART = A recetter techniquement / ECR = En cours de recette / RTE = Recetté techniquement / ECH = En cours d'homologation / HOT = Homologation terminée / ECC = En cours de correction / TER = Terminé / ABD = Abandonné

Tâches																															
Lot	Groupe	N°	Description	Commentaire	Centre / Service	Statut	Techno	RaF totale	Charge totale	RaF	Charge	Qui	Date de début	Date de fin	Réalisé	RaF	Charge	Qui	Date de début	Date de fin	Réalisé	RaF	Charge	Qui	Date de début	Date de fin	Réalisé				
Acquisiti on d'une solution		L1-01	Choix de la solution		Ratio	10 - TER		0	20	0	20	FCE	01/12/13	01/02/14	100%	0	0				0%	0	0				0%				
		Total du lot						0	20	0	20							0	0							0	0				
Installatio n / Mise en place		L2-01	Installation et mise en place		Ratio	10 - TER		0	16	0	5	FCE	01/03/14		100%	0	10	FCE FTL MBE			100%	0	1	FCE				100%			
		Total du lot						0	16	0	5							0	10							0	1				
Paramétrage Lot 1		L3-01	New CGA-Contact Migration	Intégré dans le projet New CGA-Contact	Ratio	11 - ABD		0	0	0	0	FCE ???			0%	0	0	???			0%	0	0	FCE				0%			
		L3-02	New CGA-Contact Chargement Quotidien	Intégré dans le projet New CGA-Contact	Ratio	11 - ABD		0	0	0	0	FCE ???			20%	0	0	FCE			0%	0	0	FCE				0%			
		L3-03	New CGA-Contact Fichiers Retour	Intégré dans le projet New CGA-Contact	Ratio	11 - ABD		0	0	0	0	FCE ???			25%	0	0	???			0%	0	0	FCE				0%			
		L3-01B	New CGA-Contact (Assistance)		Ratio	10 - TER		0	10	0	10	FCE			100%	0	0				0%	0	0					0%			
		L3-04	Migration Monoprix vers Kyriba	Splitage, contrôle et conversion	Ratio	10 - TER		0	5	0	1	FCE			100%	0	2,5	FCE			100%	0	1,5	FCE				100%			
		L3-05	Allemagne	Intégration du fichier GL	Ratio	10 - TER		0	11	0	2	FCE	21/03/14		100%	0	7	FCE			100%	0	2	FCE				100%			
		Total du lot						0	26	0	13							0	9,5							0	3,5				
Paramétrage Lot 2		L4-01	Remplacement SETI-Map		Ratio	3 - ECD		5,2	18	0	4	FCE			100%	2,2	11	FCE NSIT			80%	3	3	FCE				0%			
		L4-02	Outil de test de fichiers	Remplacement des tests manuels (fichiers classiques)	Ratio	1 - ECA		9,8	10	1,8	2	FCE			10%	6	6	???			0%	2	2	FCE				0%			
		L4-03	Flux Opérations Spéciales		Ratio			20	20	4	4	FCE			0%	13	13	???			0%	3	3	FCE				0%			
		L4-04	Chargement Bulk inter-applications (Ligis et Reliance)		Ratio			10,2	12	0,2	2	FCE			90%	7	7	???			0%	3	3	FCE				0%			
		L4-05	PCA Financement		Ratio	1 - ECA		8,2	10	0,2	2	FCE			90%	6	6	???			0%	2	2	FCE				0%			
		L4-06	INSEE		Ratio	1 - ECA		8,2	10	0,2	2	FCE			90%	6	6	???			0%	2	2	FCE				0%			
		L4-07	Blanchiment		Ratio	1 - ECA		5,5	6	0,5	1	FCE			90%	3,5	3,5	???			0%	1,5	1,5	FCE				0%			
		Total du lot						67,1	86	6,9	17							43,7	52,5							16,5	16,5				
Total général						67,1	148	6,9	55							43,7	72							16,5	21						

Nbre total de tâches	15
En cours d'analyse (1 - ECA)	4
Analyse terminée (2 - ATE)	0
En cours de développement (3 - ECD)	1
A recetter techniquement (4 - ART)	0
En cours de recette (5 - ECR)	0
Recetté techniquement (6 - RTE)	0
En cours d'homologation (7 - ECH)	0
Homologation terminée (8 - HOT)	0
En cours de correction (9 - ECC)	0
Terminé (10 - TER)	5
Abandonné (11 - ABD)	3

Liste des figures

Figure 1 : Principes de l'affacturage	11
Figure 2 : Organigramme CGA.....	12
Figure 3 : Organigramme DSI.....	14
Figure 4 : Magic Quadrant – Outil d'intégration de données.....	25
Figure 5 : Schéma simplifié des flux du SI CGA actuel	27
Figure 6 : Schéma simplifié des flux du SI CGA intégrant l'ETL.....	29
Figure 7 : Cas d'utilisation : Réception de fichiers clients non standards.....	32
Figure 8 : Cas d'utilisation : Réception de fichiers clients par mail	33
Figure 9 : Cas d'utilisation : Chargement inter-applications	33
Figure 10 : Cas d'utilisation : Conversions complexes avec accès à un référentiel.....	34
Figure 11 : Cas d'utilisation : Extraction des données métiers vers un fichier plat	34
Figure 12 : Cas d'utilisation : Contrôles et traitements spécifiques de fichiers reçus.....	35
Figure 13 : Diagramme d'évaluation des fournisseurs.....	37
Figure 14 : La suite Information Server	39
Figure 15 : Architecture logique Information Server	44
Figure 16 : Environnement de non-production Datastage.....	46
Figure 17 : Environnement de production Datastage.....	47
Figure 18 : Diagramme des flux Datastage	49
Figure 19 : Analyse du flux Monoprix.....	60
Figure 20 : Job SCF_Monoprix_0_Lancement (1)	61
Figure 21 : Job SCF_Monoprix_0_Lancement (2)	61
Figure 22 : Job SCF_Monoprix_1_Lecture.....	61
Figure 23 : Job SCF_Monoprix_2_Conversion (1).....	62
Figure 24 : Job SCF_Monoprix_2_Conversion (2).....	62
Figure 25 : Job SCF_Monoprix_2_Conversion (3).....	62
Figure 26 : Container fn_Control_Doublon_Fichier	63
Figure 27 : Job SCF_Monoprix_2_ListeDO.....	63
Figure 28 : Job SCF_Monoprix_3_Ecriture400.....	64
Figure 29 : Job SCF_Monoprix_4_Rejets.....	64
Figure 30 : Schéma de déclenchement du job SCF_Monoprix_0_Lancement	65
Figure 31 : Analyse de l'intégration du fichier comptable Allemand.....	67
Figure 32 : Job parallèle NAA_FichiersComptables_GL	68
Figure 33 : Job parallèle NAA_FichiersComptables_Lancement.....	69
Figure 34 : Analyse de la maquette « Mise à jour du disponible »	71
Figure 35 : Job NCG_POC_Maj_Dispo.....	71
Figure 36 : Job NCG_POC_EcritureXML.....	72
Figure 37 : Job NCG_POC_Lecture_XML.....	72
Figure 38 : Job TraitementMails_Job.....	76
Figure 39 : Job TraitementMails_Seq	77
Figure 40 : Job TraitementFichiers_Seq	78
Figure 41 : Job GEN_Transfert.....	79
Figure 42 : Utilisation du stage "Complex Flat File"	81
Figure 43 : Job SCF_CoverionMultiFormat.....	82
Figure 44 : Job DST_Template_CSV_DB2.....	87
Figure 45 : Job DST_Template_CSV_Plat400.....	88
Figure 46 : Job DST_Template_DB2_CSV.....	88
Figure 47 : Job DST_Template_DB2_SqlServer.....	88
Figure 48 : Job DST_Template_SqlServer_CSV.....	89

Liste des tableaux

Tableau 1 : Planning des tâches de la phase d'acquisition.....	45
Tableau 2 : Les droits Datastage	50
Tableau 3 : Les coefficients appliqués pour le chiffrage des flux Datastage	53
Tableau 4 : Planning des tâches de la phase mise en place de l'outil	58
Tableau 5 : Format du fichier PJ_Recus.csv	75
Tableau 6 : Format du fichier Param_Mails.csv	76
Tableau 7 : Format du fichierMailsATraiter.csv.....	76
Tableau 8 : Format du fichier Param_Fichiers.csv.....	77
Tableau 9 : Planning des tâches de la phase réalisation.....	90
Tableau 10 : Macro-Planning du projet.....	93
Tableau 11 : Bilan budgétaire du projet.....	100

Mise en place d'une solution ETL dans le cadre de la rationalisation d'un système d'information.

Mémoire d'Ingénieur C.N.A.M., Paris 2015

RESUME

La Compagnie Générale d'Affecturation connaît une croissance continue depuis plusieurs années. Pour suivre cette expansion, un schéma directeur du système d'information est mis en place. Une direction fondamentale de ce schéma est l'acquisition d'outils performants. Ce mémoire décrit le projet mené, dans ce cadre, pour choisir une solution ETL (Extract-Transform-Load) la plus adaptée au contexte, mettre en place cette solution et l'intégrer au SI de CGA. Datastage (solution IBM) devient ainsi une brique centrale du SI et centralisera l'ensemble des opérations d'intégration de données entre sources et cibles hétérogènes.

Mots clés : ETL, Datastage, schéma directeur, rationalisation. intégration des données

SUMMARY

CGA is experiencing continuous growth for several years. To follow this expansion, an information systems master plan is set up. A fundamental direction of this scheme is the acquisition of powerful tools. This document describes the project completed, in this context, to choose an ETL (Extract-Transform-Load) adapted to the context, to implement this solution and to integrate in the CGA information system. Datastage (IBM solution) becomes a central component of the information system and will centralize all data integration operations between heterogeneous sources and targets.

Keywords : ETL, Datastage, rationalization. data integration