De Datamanager à Stambia : l'ETL au cœur de la BI

Auteur : Stéphane Roussel, CEO

L'informatique, depuis ses débuts, amène son lot de prêcheurs parcourant le monde et illuminant de leur credo l'armée de développeurs qui rêvent d'embarquer auprès de ces capitaines pour un voyage initiatique vers des territoires inconnus. Si ces nouveaux conquérants ne cherchent pas tous le fabuleux métal évoqué par José-Maria de Heredia, leur quête repousse toujours plus loin les limites des connaissances digitales et leur exploitation. Beaucoup de navires s'échouent, d'autres restent à quai mais leur nombre toujours croissant permet de faire miroiter des terra incognita aux ressources inouïes. A la question de savoir d'où leur est venue cette illumination, la réponse est, pour la majorité, qu'il suffisait d'y penser tel Christophe Colomb écrasant l'œuf pour le faire tenir debout [1].

La poule et l'œuf

L'analogie avec l'œuf ne s'arrête pas là. Plus de capitaine; le terme de gourou est largement utilisé dans le secteur informatique et qui dit gourou dit disciples. Les gourous exposent les idées et les disciples en réalisent la bonne exécution. Peut-être est-ce parfois l'inverse. Datamanager, ETL acquis par IBM avec la solution Cognos en 2008, lequel l'avait lui-même acheté à Relational Matters, s'appelait alors DecisionStream et avait été inventé en 1993. Il était, selon Ralph Kimball un des deux gourous de la Business Intelligence, l'ETL le plus «dimension friendly» de l'époque. Ralph Kimball expose en 1996 les grands principes de mise en œuvre des datawarehouse dans son ouvrage « The datawarehouse toolkit », soit trois ans après la création de Relational Matters. Qui est l'œuf, qui est la poule ?

Ralph Kimball considère que le datawarehouse est une somme de datamarts, a contrario de Bill Inmon qui voit dans les datamarts une extraction du datawarehouse. La bataille des gourous est lancée, l'approche de Kimball étant maintenant largement préférée pour la rapidité de sa mise en œuvre. Il serait trop fastidieux de lister ici les 38 règles énoncées par Kimball; nous nous limiterons à celles qui nous intéressent tout particulièrement dans le cadre de cet article. Les principes en sont brièvement les suivants :

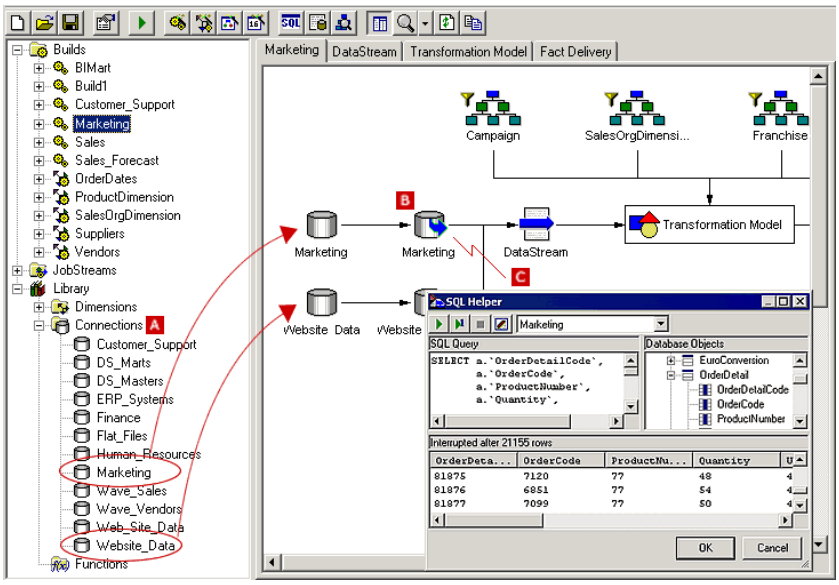
- Compréhension des besoins.
- Performance et facilité d'utilisation.
- Mise en œuvre par un cycle itératif.

Les bases de la méthode agile sont déjà présentes d'un point de vue conceptuel et, avec l'augmentation des volumes à traiter, l'approche du datawarehouse prônée par Kimball permet de s'affranchir des problématiques de temps de réponse.

[1] Anecdote attribuée aussi à son contemporain, l'architecte, sculpteur, peintre et orfèvre Filippo Brunelleschi (1377-1446).

Le datawarehouse doit, entre autres fonctionnalités, permettre :

- La hiérarchisation des informations dimensionnelles.
- La mise à disposition de dimensions communes (Conformed dimensions) pour les différents contextes métiers.
- D'historiser les données, notamment par la mise en place de SCD (Slowly Changing Dimension).
- De disposer de clés de substitution dédiées (surrogate keys).



Bingo ! Decision Stream permet de faire tout cela à travers son interface de modélisation dédiée à la création d'un datawarehouse et des datamarts associés. Il dispose d'un moteur permettant des transformations et l'agrégation des données ; on parle alors d'ETL (Extract Transform Load). Il est, selon Kimball, l'outil pour construire les datawarehouse et datamarts.

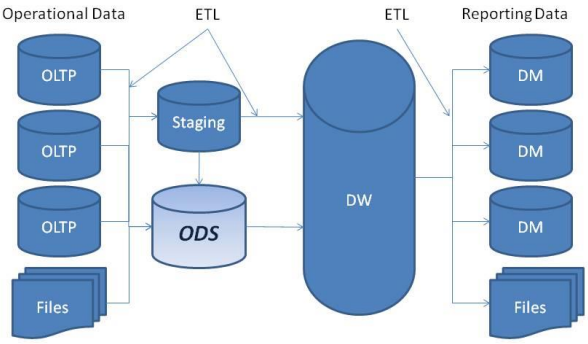
La mise en œuvre de ces principes participe à ce que l'on appelle la dénormalisation de la base de données, c'est-à-dire la violation des sacro-saintes formes normales édictant les bonnes pratiques en matière de modélisation de bases de données. Sacrilège! Une nouvelle religion est née et les prophètes des temps modernes vont arpenter le monde en évangélisant les autochtones arc-boutés sur leurs traditions ancestrales.

La Business Intelligence, l'Eldorado des années 1990

Les conquistadors font suite aux explorateurs et vont semer leur fabuleux méteil sur les territoires vierges. Portés par les alizés, des convois entiers prennent le large. BusinessObjects et Cognos en sont les fers de lance et embarquent dans leur sillage tous ceux qui veulent les rejoindre. Cognos rachète Relational Matters et DecisionStream deviendra DataManager; BusinessObjects rachète ActaWorks et renomme son ETL «BODI» (Business Objects Data Integrator). Les convois grossissent et proposent une panoplie d'outils permettant de couvrir tous les besoins en Business Intelligence. La méthodologie du datawarehouse est au cœur de ces odyssées et l'ETL est le vaisseau amiral de la flotte. Certains esprits dissidents créent leur propre flottille avec plus ou moins de succès, proposant de nouvelles approches. C'est l'explosion des solutions de Business Intelligence avec des outils front end de reporting et d'analyse, et surtout le développement de nouvelles approches en matière de stockage de données, d'architecture et d'alimentation.

Les flots grossissants, la performance devient un élément crucial et l'ETL est souvent un goulet d'étranglement dans lequel des courants toujours plus forts déversent leur flot de données. Les bases de données offrent des architectures permettant des extractions massives grâce aux performances croissantes des machines et aux évolutions récentes dans le stockage de données qui autorisent des volumétries importantes (bases de données colonne, bases de données vectorielles). Mais le nerf de la guerre repose sur les ETL qui font les navettes entre les différents composants. Là aussi on voit émerger des spécialistes proposant des voies alternatives; les ETL (Extract Transform Load) traditionnels proposent des moteurs de plus en plus efficaces pour les opérations massives mais, en raison des transformations nécessaires à la création d'un datawarehouse, cette approche, malgré les progrès techniques, conduit à des étranglements importants et les ELT (Extract Load Transform, principalement Sunopsis) se reposent sur la puissance des bases de données pour réaliser les transformations. Un flux ETL peut prendre plusieurs heures lorsqu'il s'agit de faire des agrégations et opérations sur plusieurs millions de lignes ; c'est le moteur de l'ETL qui réalise cette opération et qui représente souvent, comme dans les systèmes biologiques, le facteur limitant défini en 1840 par Justus von Liebig [2]. Dès lors, les transferts massifs de données et surtout leur transformation conduisent à des situations proches du blocage car la performance d'un moteur d'ETL ne peut atteindre celle d'un ELT. En effet, les bases de données sont totalement optimisées pour réaliser des traitements sur de gros volumes de données et l'approche alternative des ELT permet de faire effectuer ces grosses transformations par la base de données elle-même. Stambia est directement issu de cette dernière approche.

ODS within Data Warehouse Architecture



Pour ne pas surcharger les systèmes opérationnels, on crée les staging area ou ODS (Operational Data Store) qui servent ensuite à alimenter les datawarehouse. C'est lors du process d'alimentation du datawarehouse que sont créées les transformations par l'ETL. Ce process peut être exécuté par le moteur de l'ETL, qui est limité par la complexité du calcul et la volumétrie ou par la base de données , on parle alors d'ELT, offrant une meilleure performance mais nécessitant d'envoyer un ordre qui peut parfois être délicat à adresser à la base de données.

L'exploration a laissé place à l'exploitation; les concepts maîtrisés du reporting et de l'analyse de données sont maintenant largement diffusés à travers le monde. Le datawarehouse est incontournable pour qui veut proposer des tableaux de bords pertinents à l'intérieur de l'entreprise et ceux qui ne le font pas sont confrontés à des problèmes d'incohérence de données, de performance, d'ilotage des différents métiers.

[2] Justus von Liebig (1803-1873) est un chimiste allemand ayant popularisé la Loi du minimum énoncée en 1828 par Carl Sprengel (1787-1859), botaniste de même nationalité.

Les grandes découvertes des années 2000

Après les explorateurs et les conquistadors, des explorations plus scientifiques sont organisées par Bougainville, La Pérouse et Cook pour ne citer qu'eux. Ils embarquent dans leurs convois des naturalistes, botanistes, physiciens, astronomes et géographes poussés par le souffle des Lumières. L'esprit des sciences emboîte le pas à la colonisation et cette curiosité est à l'aube de grandes découvertes. Il ne s'agit plus d'imposer sa présence à tout prix mais d'enrichir ses connaissances par l'étude de ces nouvelles contrées. Ces «savanturiers», parfois un peu naïfs, ne survivront pour la plupart pas à leurs périples mais leurs découvertes permettront des avancées scientifiques majeures dans les domaines de l'astronomie, la biologie et la médecine notamment.

Les années 2000 voient la concentration des éditeurs de Business Intelligence absorbés par les géants de l'informatique. Business Objects est racheté par SAP; Cognos par IBM; Siebel par Oracle. Pour ces grands éditeurs il ne s'agit pas de faire de nouvelles découvertes mais dans un premier temps d'intégrer ces nouvelles acquisitions dans leur large gamme de produits. Place au marketing; les scientifiques attendront ! Mais ces derniers n'attendent pas et l'émergence de nouvelles technologies est un formidable moteur-fusée pour ces esprits en ébullition. Internet explose et avec lui de nouvelles possibilités d'échanges de données sont offertes : le SOA (Service Oriented Architecture). Les concepts ne sont pas complètement nouveaux mais ces solutions deviennent exploitables à grande échelle avec notamment l'utilisation des web services. C'est également durant cette période que le brillant et humble Doug Cutting crée Hadoop, nommé ainsi en clin d'œil au doudou de son fils.

Ces deux apports sont à l'aube de révolutions majeures qui vont bouleverser le paysage d'une informatique établie et concourent à développer le passage d'un système centralisé à un système décentralisé. Les aspects principaux de cette évolution latente sont :

- Déporter les applications :

  Internet permet dorénavant de disposer d'applications d'entreprise externes ou d'utiliser des applications SAAS (software as a service). Ces applications sont toutefois reliées aux applications d'entreprise et des échanges inter-applicatifs deviennent nécessaires. Les ETL traditionnels orientés données et traitements par flux sont secondés par des outils EAI et ESB pouvant gérer des échanges multiformes au fil de l'eau.

- Ouvrir les sources sur des données externes non propriétaires :

  On assiste au développement des applications sur le cloud mais apparaissent en outre de nouvelles sources de données externes à l'entreprise avec le développement d'Internet, du GPS, des réseaux téléphoniques mobiles, des réseaux sociaux, etc. Ces dernières sources amènent de nouvelles contraintes qui conduisent au développement de Hadoop car les systèmes traditionnels ne peuvent répondre aux 3 V (Volumétrie, Vélocité, Variété) et Yahoo en est le précurseur. Les données sont de plus en plus volumineuses et hétérogènes ; les intégrer pour en extraire la substantifique moelle[3] revient à résoudre la quadrature du cercle.

- Repenser l'organisation :

[3] « C'est pourquoi fault ouvrir le livre et soigneusement peser ce que y est déduict. [...] Puis, par curieuse leçon et meditation frequente, rompre l'os, et sugcer la substantificque moelle, [...]. » François Rabelais, *La vie très horrifique du grand Gargantua, père de Pantagruel, jadis composée par M. Alcofribas abstracteur de quintessence. Livre plein de Pantagruélisme, 1534.*

Les directions informatiques deviennent les clients des métiers internes. En 2015, Gartner estimait que 37% des projets informatiques n'étaient pas pilotés par les DSI. Les métiers financent ainsi et organisent à leur gré des projets sans interaction avec la DSI. Le datawarehouse a conduit à la dé-normalisation, le monde digital amène la dés-organisation. D'où l'impression confuse d'un capharnaüm difficilement compréhensible et surtout une absence de vision globale d'entreprise qui s'en dégage. On parle de méthode agile pour répondre le plus rapidement possible aux besoins métiers qui nécessitent une réponse rapide à un monde en perpétuelle évolution.

C'en est donc fini du datawarehouse ? Non car il a prouvé son efficacité et il demeure la mémoire de l'entreprise mais il doit s'ouvrir pour proposer de nouveaux modes d'intégration afin de prendre en compte l'évolution galopante.

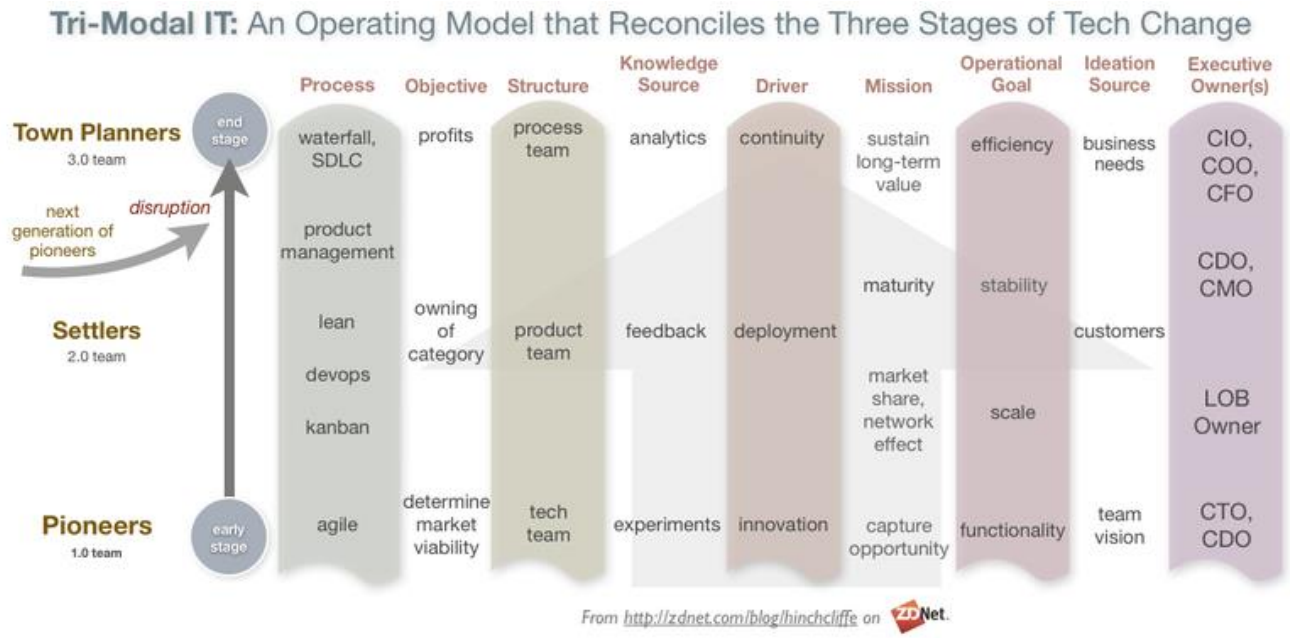
Physique quantique et relativité générale

Mais comment combiner ces deux approches qui semblent tellement opposées ? Comme en physique aurait-on deux modèles antinomiques apparemment impossibles à rapprocher, la physique traditionnelle avec la théorie de la relativité générale et la physique quantique issue des modèles probabilistiques? En effet, si l'approche datawarehouse est en phase de maturité, l'approche Big Data issue de Hadoop peut sembler incompatible tant elle véhicule des particules incontrôlables et pourtant toutes deux rejoignent le même but : mieux comprendre et analyser les données qui gravitent autour de nous. Avec une petite différence toutefois. L'approche Datawarehouse est plus tournée vers l'analyse tandis que celle du Big Data propose des modèles prédictifs. Ce qui revient à dire en termes simples, permet dans le premier cas de comprendre ce qui est arrivé pour imaginer ce qui va se passer dans le second. Passé contre futur vous avez dit ? Flaubert l'exprimait ainsi : « L'avenir nous tourmente, le passé nous retient, c'est pour ça que le présent nous échappe » [4].

Et c'est bien là où le bât blesse. Car nos voyageurs temporels, après avoir parcouru les mers, ont maintenant les yeux rivés aux étoiles et, si leurs vaisseaux possèdent des voiles, celles-ci les gênent pour voler. Les ETL ont certes connu des évolutions, se sont perfectionnés mais, pour échanger des données avec des web services, leur architecture traditionnelle rend l'opération ardue et fastidieuse car ils se doivent d'incorporer la technologie à chaque nouvelle connexion. Les protocoles développés par les EAI et ESB semblent être incontournables car ils permettent de fluidifier les échanges et surtout amènent l'agilité qui manque aux ETL.

Dès lors, que faire ? Rester à quai en regardant les uns voguer sur les flots et les autres rejoindre les étoiles ? La difficulté de répondre à cette dualité porte un nom : l'informatique bimodale selon le très respecté Gartner. Des esprits avisés comme Simon Wardley ont commencé à élaborer une stratégie de migration permettant le passage de l'un à l'autre mode. L'informatique bimodale n'est pas encore établie qu'on parle déjà d'informatique trimodale. Nos explorateurs sont, comme dans les récits sur la découverte des Amériques , chronologiquement des « pionniers » puis des « colons » et enfin des « urbanistes » comme l'illustre notre schéma.

[4] In Lettres à Louise Collet de Gustave Flaubert (1821-1880), romancier du courant réaliste français.

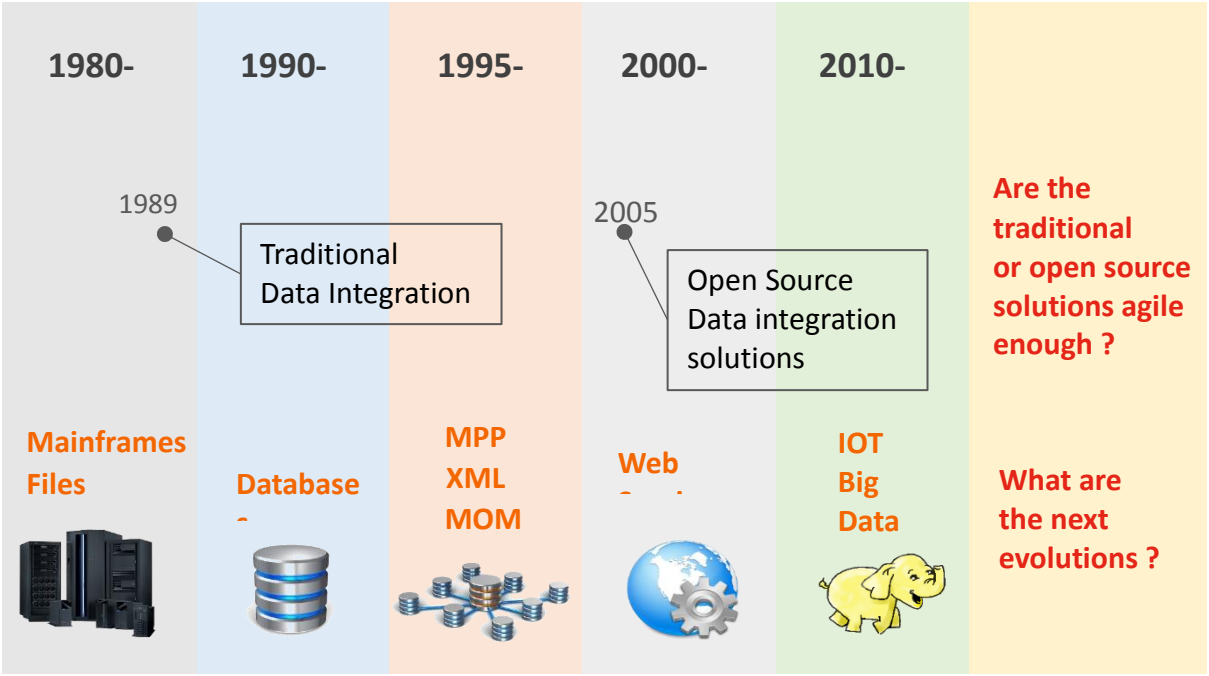


Cependant comment passer de Santa Maria à l'USS Enterprise sans dommages? Faut-il associer à chaque étape des outils et architectures différents au risque de multiplier compétences, matériels et équipages ? Sunopsis avait ouvert la voie en proposant une approche déportée pour le traitement des flux, l'ELT. Stambia, dans la continuité de cette philosophie, permet non seulement de faire exécuter les traitements par les bases de données mais ouvre aussi la voie à d'autres technologies comme les web services, le XML, les données non structurées, les IOT (Internet Of Things), etc. Bref, vous l'avez compris, à environnement multiforme, outil multiforme. Le lecteur ayant lui-même dompté de nombreux océans n'en est pas à son galop d'essai et rétorquera avec une dubitativité nourrie d'expérience, qu'un outil qui peut tout faire ne fait rien de bien ! A nouveau paradigme, nouvelle conception : Stambia propose une méthodologie pilotée par les modèles donc très orientée métier et réussit le tour de force de faire des flux beaucoup plus rapidement que les ETL classiques par la délégation des transformations, en s'appuyant sur une plateforme éprouvée et reconnue : Eclipse. Par cette approche, l'outil peut évoluer très rapidement car l'intégration de nouvelles technologies en est facilitée et, alors que l'on pouvait se permettre en voguant sur les flots de scruter longuement l'horizon avec sa longue-vue, la marche vers les étoiles requiert des évolutions rapides pour éviter comètes et pluies d'astéroïdes.

A la conquête de nouveaux mondes

2015 annonce une nouvelle époque avec la fermeture du Ralph Kimball Group. De nouveaux challenges apparaissent et la DSI ne peut pas tout orchestrer. Les métiers prennent la main et ont besoin d'outils agiles et maîtrisables par des non-initiés pour s'adapter aux exigences d'un monde en perpétuelle évolution. Datamanager était un prince dans la construction des datawarehouse ; cependant l'augmentation des volumes et des types d'échanges inter-applicatifs montre les limites de cette approche universelle qui ne peut répondre seule aux nouveaux enjeux.

Les choix faits maintenant par les DSI auront une incidence certaine sur les possibilités offertes demain et ces choix d'outils peuvent avoir un impact stratégique non négligeable or, qui ne sait pas vers quel port se diriger n'a pas de vent favorable [5].

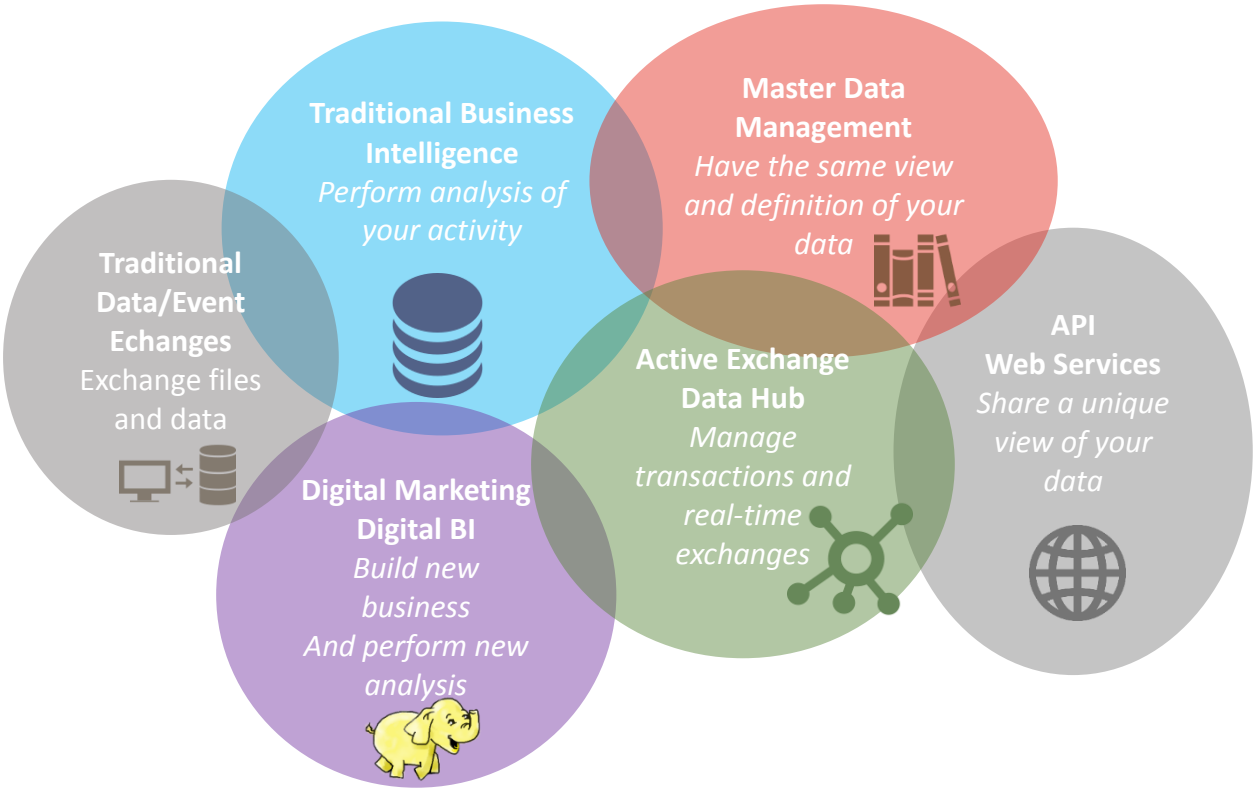


A l'augmentation des volumes, l'approche ETL de Datamanager avec un moteur intégré peut poser problème avec saturation de la mémoire et bloquer de fait les flux les plus volumineux. On peut cependant contourner ce problème en insérant directement le SQL dans des jobstreams et ainsi faire fonctionner datamanager comme un ELT en y perdant toutefois l'approche tant vantée par Kimball.

Malheureusement, en ce qui concerne les possibilités de connexion aux web services, aux IOT, à Hadoop, etc. Datamanager reste muet et regarde ces ODNI (Objets Digitaux Non Identifiés) avec l'impuissance d'un mousse ébloui par les étoiles filantes qu'il ne peut approcher. Un autre outil s'impose donc.

L'approche est délicate car la décision de choisir un outil multiforme peut conduire à un gigantesque naufrage si la démarche n'est pas maîtrisée. Les nombreux écueils sont autant d'icebergs qui affleurent sans montrer l'importance du danger latent car, au-delà des choix technologiques pris, la vision même de l'entreprise est au cœur du problème. En effet, l'anticipation des projets futurs permettra de déterminer s'il est pertinent ou non d'engager une refonte des systèmes existants. Doit-on se tourner vers le Big Data ? Les échanges de données seront-ils uniquement basés sur des ordres SQL ou doit-on prendre en compte des architecture SOA ? Y a-t-il besoin de transferts de données au fil de l'eau ? Etc. Toutes ces questions ne sont pas que du ressort de la DSI mais impliquent également les différents services de l'organisation, ils sont déterminés par la stratégie de l'entreprise sur les échanges de données entre les différents acteurs.

[5] « Ignoranti quem portum petat nullus suus ventus est » in Lettre 71 à Lucillius de Sénèque (1 av. JC à 4 ap. JC – 65 ap. JC) , philosope, dramaturge et homme d'état romain.



Le choix d'une migration doit donc s'inscrire dans un schéma directeur car il n'est pas neutre financièrement même si Stambia propose des outils permettant d'en optimiser les coûts avec, comparativement, un coût de mise en œuvre de l'ordre de 30% comparé à celui des solutions alternatives.

Revenons sur Terre !

Supposons qu'après plusieurs analyses de besoins, revues de process, POC, évaluations des coûts de licences, la solution Stambia soit approuvée comme outil d'échanges de données dans l'entreprise. La question se pose alors d'une migration sans régression et sans coût exorbitant depuis DataManager.

Pour pallier ces problèmes, un outil d'aide à la migration est proposé par Stambia et permet de reprendre l'existant avec un minimum d'effort, que Datamanager utilise les builds ou bien uniquement les job streams en mode SQL. Les fonctions et variables définies dans Datamanager peuvent également être reprises de manière unitaire ou plus globale, à l'aide de templates dans Stambia permettant leur réutilisation et industrialisation.

BI2B a pu mettre en œuvre les deux approches (builds et job streams) et la souplesse de Stambia permet de rendre une copie très propre avec de surcroît des gains de performance non négligeables. Les coûts très maîtrisés ainsi que les gains d'exploitation rendent même l'opération rentable à 2-3 ans, d'après les clients chez qui l'opération de migration a été réalisée.

Notre retour d'expérience sur le sujet permet d'affirmer que le temps de migration est fortement réduit en utilisant les outils de migration proposés par Stambia par rapport à d'autres solutions. En se basant sur les POC que nous avons réalisés sur le sujet, il apparaît que le gain est de 50 à 70% par rapport à une refonte manuelle unitaire des flux. Le gain est également important en exploitation car

la simplicité d'utilisation de Stambia, ainsi que son approche dirigée par les modèles, amène des gains de l'ordre de 25% en comparaison d'un ETL classique.

BI2B a déjà réalisé trois migrations vers Stambia, toutes couronnées de succès, et qui nous renforcent dans les convictions suivantes :

- La non régression est totale lors des projets de migration.
- La limitation des temps de migration dégage du temps pour des optimisations et refontes.
- L'appropriation par le client est très bonne et, conforté par un projet de migration couronné de succès, ouvre la voie vers de nouvelles fonctionnalités (XML, CDC, MDM, etc.).
- Il y a toujours des gains de performance sur les flux migrés.
- Limitée au simple périmètre de la migration, sans prendre en compte aucune autre amélioration ou fonctionnalité, l'opération est rentable en intégrant les coûts de licences dont le modèle simple et visible rassure, ainsi que grâce aux gains d'exploitations une fois la migration achevée. Nous parlerions en économie d'une optimisation du coût d'opportunité, du temps productif étant potentiellement dégagé par ces migrations.

A propos de BI2B

BI2B, par sa société BSL Consulting, intervient depuis 2001 dans le domaine de la Business Intelligence où elle a acquis une réputation d'expertise et de fiabilité. A travers ses sociétés basées en France, BSL Consulting, et en Suisse, BI2B Suisse, elle propose des présences locales aguerries aux problématiques et permettant un suivi de proximité sur les projets de ses clients.

BI2B, forte de son équipe d'experts Métiers et Technologies certifiés, de ses seize ans d'expériences positives en gestion de projets et de ses proches partenariats avec les éditeurs, vous accompagne durant tout le processus de changement et crée pour vous les solutions optimales, pérennes, évolutives et aisément appropriables que vous souhaitez . Ecoute, suggestions, prise en compte des contraintes clients, l'équipe de BI2B se veut avant tout pragmatique et propose des solutions réalistes avec une démarche itérative et d'appropriation des technologies.

Créateur et distributeur de logiciels, BI2B a conclu des partenariats forts avec Stambia -Gold Partner-, IBM Cognos -Value Advantage Plus-, ReportOne et SAS.

Les logiciels conçus par BI2B sont **BSL Security Manager** et **COMI** qui permettent respectivement :

BSL Security Manager : optimisation et délégation de la sécurité sous Cognos ;

COMI : cartographie et supervision des flux métiers avec prise en main possible par les métiers ;

Nous vous invitons à les découvrir lors de nos prochains webinars gratuits.

3 Place Mendès France
Elancourt, France
+33 1 30 66 31 08

Quai du Seujet, 10
Genève, Suisse
+41 22 566 74 63