ETL 2.0 - Eléments de compréhension

Contexte

Depuis 20 ans, les systèmes d'informations ont utilisé des solutions d'intégration de données pour déployer des initiatives nécessitant des échanges inter-applicatifs
- Entrepôts de données
- Fusion / Migration de systèmes
- Compliance
- etc

Ces dernières années, les suites d'intégration de données se sont enrichies de nombreux modules, destinés à répondre aux nouvelles initiatives stratégiques des entreprises
- Déploiement de référentiels unifiés
- Vision 360° des clients
- Amélioration de la Supply-chain
- Mise en œuvre d'interfaces B2B normalisées
- Traçabilité des circuits de l'information
- Mise à disposition d'informations décisionnelles
- Data Analytics
- Compliance
- etc

Aujourd'hui, on parle de plus en plus de nouvelles initiatives autour du Big Data, qui consiste à préparer le traitement des données nouvelles ou enfouies afin de dégager toujours plus de valeur autour à partir de ses informations : Les Big Data.

Seulement voilà, les solutions d'intégration de données standards du marché ne suivent pas, ni du point de vue de la volumétrie, ni du point de vue des attentes métiers. Sur tous les projets actuels, les équipes d'études ont dû trouver des moyens de pallier aux performances des ETL standards afin de délivrer les résultats en accord avec les besoins métiers : Expertise, ajout de matériel, déport des traitements dans les bases de données, autant de "bonnes pratiques" qui entraînent des augmentations significatives de coûts de possessions et qui impactent l'agilité des projets.

Alors que les projets demandent toujours plus de ressources, le contexte actuel de crise économique amène des DAF et les DSI à chercher des angles pour réduire les coûts de possession des applications.

On peut poser la question : Comment ces moteurs d'intégrations, inadaptés aux projets d'aujourd'hui tant en terme de coût que de capacité de traitement, vont-ils se comporter face à des volumes 44 fois supérieurs (estimation du Gartner d'augmentation des volumétries entre 2009 et 2015) ? Lorsqu'on voit le coût des infrastructures actuelles destinées à intégrer des données opérationnelles en BATCH nocturnes pour quelques dizaines d'utilisateurs, quel sera le coût des mêmes solutions

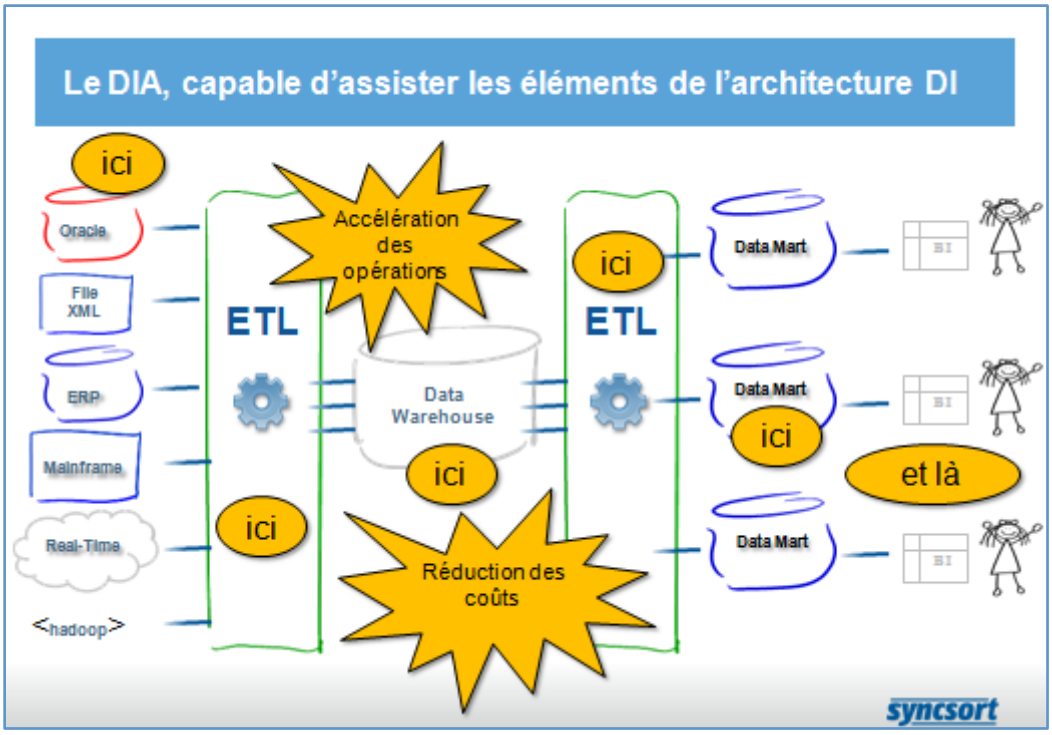
pour traiter au fil de l'eau des données opérationnelles ET externes (terminaux mobiles, réseaux sociaux, ...) en quasi temps-réel et servies à des milliers de personnes ?

Pour Syncsort, la réponse est évidente : Sans aide, les architectures d'intégration de données actuelles vont échouer sur les projets d'aujourd'hui dont la volumétrie va exploser, et ne seront pas déployées pour les projets de demain car d'emblée jugées trop coûteuses.

ETL 2.0 - Proposition de valeur

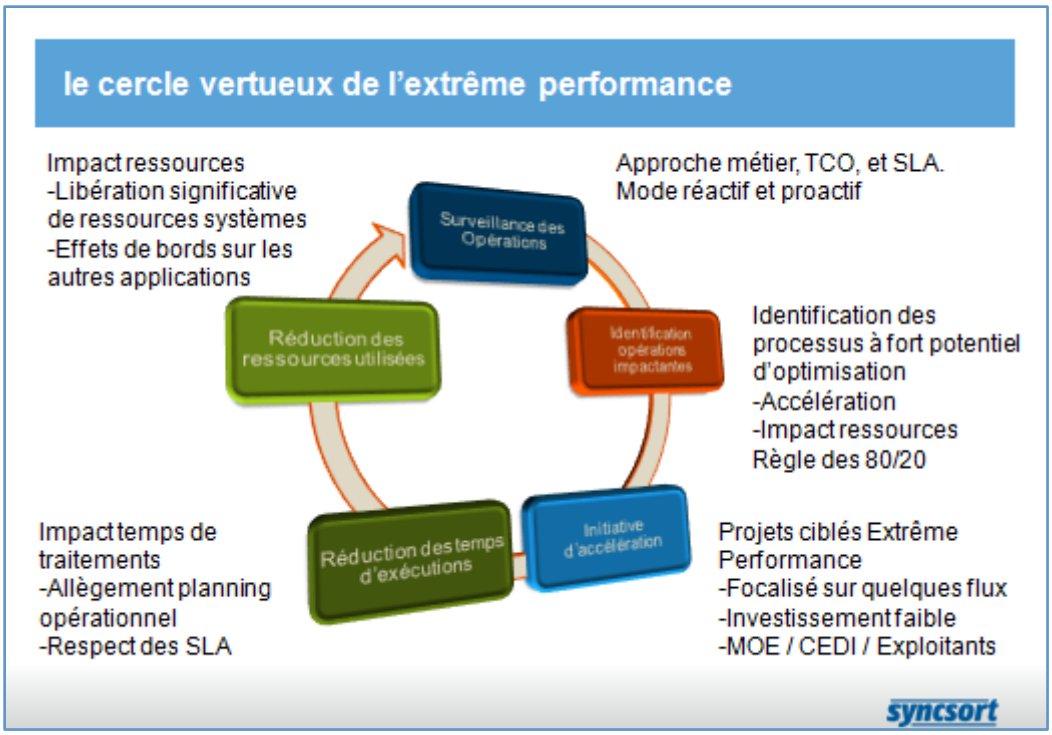
Pour débloquer la valeur des projets d'intégration de données, il va falloir significativement améliorer les performances des couches d'intégration de données, et dans le même temps en réduire les coûts de possession, et ce en renforçant les règles de développements urbanisés, afin de rester (ou de redevenir) réactif en alignement avec les besoins métiers.

Cette proposition de valeur, la stratégie ETL 2.0 la matérialise par la mise à disposition transverse d'un élément d'architecture remarquable par ses performances, son comportement vis-à-vis des ressources systèmes, et sa flexibilité. Au services des autres éléments de l'architecture d'intégration de données, ce composant appelé "DI Accelerator" (ou DIA) est destiné à en améliorer significativement les capacités de traitements ainsi qu'à en réduire les coûts de possessions.



Améliorer les capacités de traitements et réduire les besoins en ressources
La stratégie ETL 2.0 met à disposition des outils de transformation de données un "booster" pour réaliser des opérations basiques à forte volumétrie.

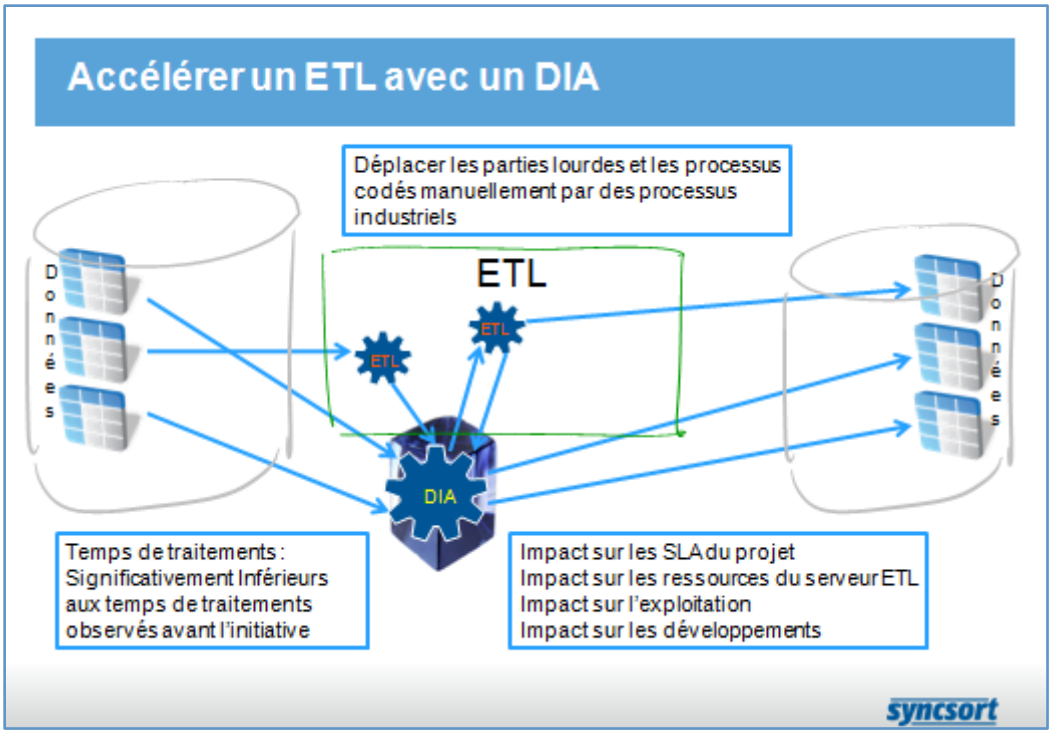
Jointures, agrégats, filtres, tris, une fois confiés au DIA, s'exécutent significativement plus rapidement et nécessitent moins de ressources pour délivrer les mêmes résultats. Cette approche, reproduite de manière cyclique, est qualifiée de "cercle vertueux de l'extrême performance". Non seulement elle réduit les temps d'exécution des processus d'intégration de données, mais de plus elle libère des ressources physiques pour l'ensemble des outils opérés sur les plates-formes. Ce double effet accélérateur est aligné avec les initiatives de contrôle / réduction des ressources physiques du système d'information.



Bénéfices d' initiatives ETL 2.0

Les bénéfices majeurs d'une initiative ETL 2.0 sont :
- Accélérer les interfaces inter-applicatives,
- Réduire les ressources requises par les interfaces inter-applicatives.
Ces deux axes de bénéfices peuvent s'appliquer pour atteindre les objectifs de différents types d'initiatives.

Ramener un projet dans ses SLA
Lorsque des interfaces inter-applicatives débordent des fenêtres de chargement qui leur ont été allouées, les effets de bords peuvent être multiples, de la non disponibilité du service aux utilisateurs jusqu'à la mise en danger de la production. L'existence d'un module accélérateur dans l'infrastructure permet de rapidement et durablement réduire les temps d'exécution des interfaces en maintenant le niveau d'industrialisation (et donc d'agilité, d'exploitabilité, et de maintenabilité) du projet concerné.

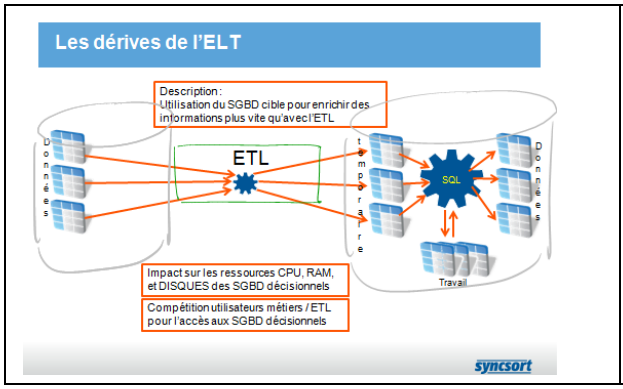


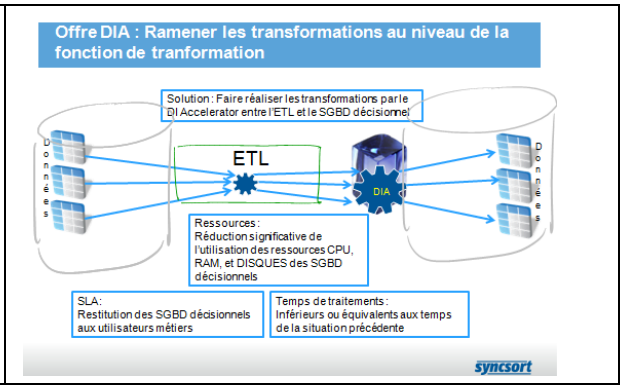
Rendre de l'agilité à une application

Le besoin de ramener un projet dans ses SLA n'est pas nouveau. Dans notre article précédent, nous avons présenté les modes opératoires communément adoptés dans ces circonstances. Parmi eux, le recours à des experts est largement répandu. Suite à de telles démarches, des interfaces peuvent avoir été "optimisées" soit par des méthodes expertes liées aux outils en place (utilisation de techniques d'experts au sein des outils déployés), soit par l'introduction d'étapes soit en code spécifique (shell script, C/C++, librairies java, etc). Ces interventions impactent l'agilité, l'exploitabilité, et la maintenabilité du projet. Elles ont pour effet d'allonger ses évolutions futures, et d'y imposer un niveau élevé d'expertise (impact sur les coûts de développement).
Déplacer ces "améliorations" sur le DIA permet de ramener le projet dans ses temps d'évolutions et ses besoins d'expertises initiaux.

Réduire ou contrôler les ressources d'un SGBD

Devant les "bonnes pratiques" de déport des transformations dans les SGBDR adoptées par les clients, les éditeurs de solutions d'intégration de données ont proposé des modules dits "push-down" dans leurs offres logicielles, pérennisant ces comportements. L'impact sur les courbes d'évolutions des ressources des SGBD s'est vite fait sentir, et de nombreux clients se retrouvent pris au "piège de l'ELT" avec des besoins d'évolutions tendant à devenir annuels, engloutissant les efforts de rationalisation des coûts faits par ailleurs.
Déplacer tout ou partie de ces transformations sur le DIA réduit significativement le besoin en ressources des SGBD ainsi déchargés, et permet d'en réduire - voir d'en inverser - la courbe d'évolution.





Monter une plate-forme "Big Data"

Comment imaginer que des solutions d'intégration de données qui ont montré leurs limites sur des chargements périodiques de données opérationnelles sont adaptées au traitement de données disparates et significativement plus volumineuses en quasi temps-réel et servies à un nombre important d'utilisateurs ?

Alors que les architectures de références se cherchent encore, il est pertinent de s'attendre à l'émergence de nouvelles solutions adaptées aux caractéristiques des Big Data, tant en matière de stockage que d'implémentation de logiques fonctionnelles. Ces solutions devront se montrer beaucoup plus rapides et moins coûteuses que les plates-formes standards du marché. Beaucoup plus rapides car les volumes de données seront nettement supérieurs à ceux sur lesquels les solutions standards échouent aujourd'hui et beaucoup moins coûteuses afin de permettre l'accès à l'analyse des Big Data à toutes les entreprises du marché.

L'approche ETL 2.0 s'appuie sur des éléments matériels et logiciels répondants à ces caractéristiques. Elle s'applique à revisiter les fonctions fondatrices des couches d'intégrations de données pour les rendre plus performantes, moins consommatrices, et moins coûteuses à acquérir et à posséder.

ETL 2.0 - composants fondateurs

Afin de délivrer des performances supérieures en réduisant l'utilisation des ressources, la stratégie ETL 2.0 s'appuie sur des éléments significativement

- plus efficaces
- moins consommateurs
- plus flexibles
- moins chers

que les solutions standards du marché de l'intégration de données.

De tels éléments existent-ils ? On est en droit de se poser la question, alors que les technologies d'intégration de données semblent être matures (et donc, par déduction, optimisées).

La performance, une affaire de spécialiste

Au fil des 10 dernières années, les éditeurs de solutions ETL ont enrichi leurs solutions de modules, fonctionnalités, capacités complémentaires afin de pouvoir se positionner sur les initiatives de leurs clients. Dès 2004 le Gartner - en bon visionnaire du marché - a modifié l'intitulé de son cadran

magique de "logiciels ETL" à "solutions d'intégration de données" en accord avec cette tendance. Seulement voilà, ces évolutions se sont faites à grand renfort d'acquisitions et d'outsourcing (largement "offshorisé"). Ces deux vecteurs d'évolutions ont fortement impacté les performances des solutions proposées. Les équipes d'engineering ont dû se concentrer sur l'intégration des solutions disparates juxtaposées pour les transformer en "plates-formes" intégrées, au détriment de l'optimisation et du contrôle des fonctions de transformation de base (tri, jointure, agrégat, I/O). C'est pourquoi ces géants, capables de couvrir des besoins fonctionnels aussi disparates que le MDM et la SOA, ont les pieds posés sur un socle d'argile vieillissant, non maîtrisé, et incapable de délivrer des performances alignées avec les volumes et enjeux d'aujourd'hui et de demain.
En rupture avec ce comportement, Syncsort s'est concentré depuis 43 ans à mettre à disposition de ses clients le moteur de transformation de données le plus optimisé possible. Société privée, Syncsort n'a jamais dérogé à sa proposition de valeur fondatrice : l'extrême performance.

Amélioré à chaque version, exploitant les nouvelles capacités matérielles, Syncsort DM Express (le moteur de transformation de Syncsort) surprend tout d'abord par la taille de son archive : 80 MO, alors que la plupart des solutions du marché dépassent les 4,7 GO des DVD classiques.

Le deuxième effet Keith Kohl de DMX (NDA : Keith Kohl, qui fut Product Manager InfoSphere DataStage pendant plus de 10 ans, est devenu Product Manager DM Express en janvier 2010) se produit lorsqu'on l'installe, car l'opération prend au plus 10 minutes, configuration comprise.

Enfin, quelques tests sur des opérations de bases à forte volumétrie finiront de vous convaincre que "oui, il existe une solution logicielle capable de réaliser des opérations d'intégrations de données de type ETL beaucoup plus vite et en consommant significativement moins de ressources que les solutions traditionnelles". Jointures, agrégats, tris, filtrages et formatages, validations techniques et fonctionnelles, autant d'opérations simples et fondatrices de l'intégration de données que Syncsort DM Express réalise sur de fortes volumétries 3X à 10X plus vite que ses précurseurs en utilisant 2X à10X moins de ressources physiques.

Le terme "précurseurs" est d'ailleurs inadapté. DM Express s'appuie sur les bonnes pratiques en vigueur dans les années soixante. A cette époque où toute l'informatique est hébergée sur LE mainframe de l'entreprise, les développeurs s'appliquent à respecter des principes forts
- Algorithmes : Chaque cycle CPU coûte de l'argent,
- Stockage : La mémoire (vive ou de masse) est une ressource rare et partagée,
- Comportement citoyen : Un ordinateur est un environnement d'exécution de plusieurs applications.
Ces principes ont guidé les équipes de développement de Syncsort depuis 43 ans.

Matériel haut de gamme ou matériel low-cost ?
Une question qui a animé bien des discussions autour de nombreuses machines à café sans jamais donner tort ou raison aux uns ou aux autres : les deux approches perdurent et ont chacune leurs champions. Récapitulons rapidement les principaux différentiateurs de ces 2 approches.

	Matériel haut de gamme	Matériel low-cost
Coûts	millions d'euros	10aines de milliers d'euros
Puissance	Puissance extrême basée sur de forts investissements en R&D	Puissance moyenne, matériel assimilé à du consommable
Fiabilité	Fiabilité maximale, pannes quasi inexistantes	Fiabilité moyenne, pouvant être compensée par de la redondance (cluster, grid, cloud)
Partitionnement	Plusieurs partitions statiques ou dynamiques par machine	Unités physiques (en général 1 blade) pouvant être "amalgamées" en unités logiques
Consommation	Orientation Green	En fonction des constructeurs
Positionnement	Socle d'applications stratégiques	Assimilé "consommable"

La préconisation ETL 2.0 : Positionner le DIA sur du matériel low-cost. Cette préconisation est justifiée par les éléments suivants

Positionnement : Si les interfaces inter-applicatives sont souvent au service d'applications stratégiques, elles sont bien souvent considérées comme de la "tuyauterie". Le niveau d'exigence les concernant est en règle générale d'assurer le transfert de données dans des délais établis.
Les projets mettent en place des procédures de reprise sur erreur, et s'appliquent en collaboration avec les exploitants à aménager des fenêtres d'exécutions permettant aux reprises d'avoir lieu. Des blades en clusters garantissent la capacité de reprise en cas de panne matérielle.
Bien sûr, les demandes métiers tendent à demander de l'information en quasi temps-réel. Les ruptures de services ne seraient alors plus tolérées. Les évolutions en matière d'architecture virtualisée et de cloud computing permettent dès à présent une garantie de service à la hauteur de cette demande.

Performances : Les benchmarks de performance démontrent que les processeurs hauts de gammes offrent des capacités de traitements significativement supérieures aux processeurs des infrastructures low-cost. Cependant, les plates-formes en production font tourner des applications et non des benchmarks. Or, comme Syncsort, de nombreux éditeurs de logiciels ont fait la constatation suivante : à nombre de cœurs équivalents, les logiciels affichent des performances équivalentes, qu'ils soient exécutés sur des plates-formes hauts de gammes ou sur des plates-formes low-cost.
De l'avis de certains, les performances seraient même meilleures sur des blades linux comparées à des machines massivement parallèles.

Coûts de possessions : 10X à 100X inférieurs sur plates-formes low-cost, en alignement avec la proposition de valeur ETL 2.0.

Cette préconisation est destinée à optimiser le rapport coûts/performances, facteur essentiel de la valeur du DIA sur le système d'informations. Il est bien sûr possible de déployer le DIA sur des serveurs hauts de gammes. Cette situation peut même être recommandée en fonction des situations.
Si le DIA est destiné à baisser le coût global du décisionnel d'une entreprise, en permettant de

ramener des transformations poussées dans le datawarehouse sur le composant ETL afin de réduire significativement l'évolution de ressources de celui-ci, positionner le DIA sur le serveur hébergeant l'ETL est plus que pertinent. Au fil des initiatives d'accélération, non seulement les flux de données seront accélérées, mais le niveau d'utilisation des ressources va lui aussi baisser. C'est ce que nous appelons le cercle vertueux de l'extrême performance.

Le cercle vertueux de l'extrême performance

La première pierre

La première brique d'une initiative ETL 2.0 sera ciblée sur un projet pilote au cours duquel la solution sera intégrée dans le système d'information, tant du point de vue technique, que fonctionnel et administratif. Des éléments tels que la prise en main par les exploitants ou la formation des études seront alors mis en place.

Tiré d'une pré-étude ayant identifié un projet pour lequel une valeur en relation avec l'investissement que cette initiative représente, ce pilote sera à même de mieux permettre à un centre d'expertise d'évaluer les situations génératrices de valeur, par exemple :

Opération	Valeur attendue	Cas client
Accélération d'un processus ETL	Réduction significative des temps de traitement. En général appliqué sur des entrepôts de données qui n'arrivent plus à terminer les chargements avant l'heure d'ouverture des services aux utilisateurs.	Zion Bank, une banque américaine, a réduit par 4 le temps de chargement de son datamart de notation des demandeurs de crédits. Revenue dans ses SLA, l'application a permis à Zion Bank de continuer à répondre aux demandeurs de crédits dans la journée, comme indiqué dans ses messages marketing.
Déchargement de transformations poussées dans les bases de données sources et cibles	Réduction significative des besoins en ressources des SGBD. Les coûts exponentiels des bases de données sont le signe de la pertinence d'une initiative dans ce domaine. Lorsque les transformations sont massivement poussées dans les bases de données, elles génèrent une augmentation significative des ressources nécessaires au bon fonctionnement des interfaces. Or la majeure partie de ces transformations peut être confiée au DIA avec des temps de traitements équivalents à inférieurs. Migrer ces traitements hors des bases de données réduit significativement les	e-Dialog est une société américaine de génération de campagnes marketings à valeur ajoutée. Les clients demandent des extractions multi-critères, et e-Dialog fournit la liste de publipostage. L'application entièrement développée en PL/SQL tendait à exploser du point de vue des temps de traitements et des ressources requises par la base de données. Après avoir migré la logique fonctionnelle sur un DIA

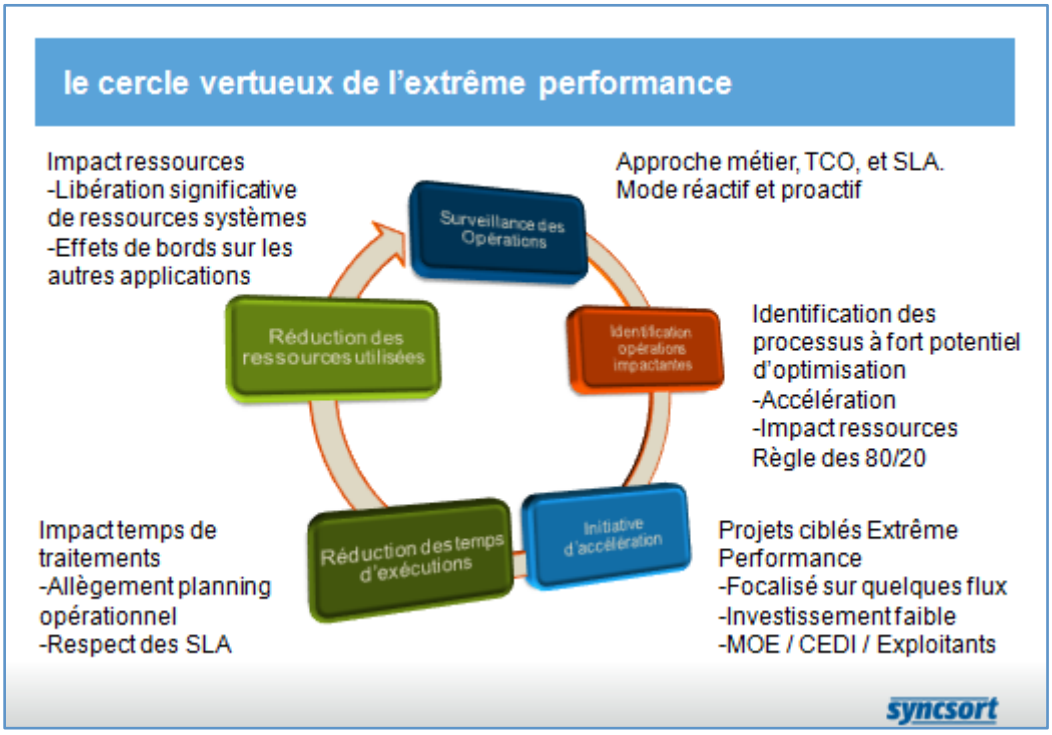
	besoins en ressources de celles-ci et permet d'en réduire les coûts de possessions	externe, e-Dialog a pu constater les axes de ROI suivants : - Réduction de 4h à 25mn du temps moyen de réponse à une demande de campagne - Réduction significative des ressources CPU, mémoire, et disque requises par la base de données - Amélioration drastique de la maintenabilité de l'application La migration de l'application a fait l'objet d'un projet qui a duré 1 mois, du démarrage au passage en production
Filtrage de journaux en amont des applications consommatrices	DM Express fait des merveilles lorsqu'il s'agit de travailler avec les disques durs. Ainsi des processus chargés de tirer de la valeur d'informations contenues dans des fichiers pourront-ils se voir fortement accélérés si le DIA est positionné en amont. En ne fournissant à l'application que les lignes qu'elle utilisera pour produire ses informations, le DIA prend à sa charge le travail d'épuration que l'application aurait sinon pris à son compte. Un cas d'usage courant est le filtrage de journaux de serveurs web en vue d'analyse comportementale. Un autre cas est le filtrage de journaux applicatifs en amont de la supervision.	Compass Bank, une banque américaine, reçoit chaque nuit 10 fichiers contenant des opérations de banques de détails (1 fichier par région). L'accroissement des volumes d'opérations avait porté à 6 heures le temps de traitement moyen observé par fichier. Non seulement l'entrepôt de données n'ouvrait plus en matinée, mais de plus certains jours d'opérations n'arrivaient pas à être chargés dans la journée suivante. Le DIA, positionné en filtre sur les fichiers avant le processus ETL, a permis de réduire à moins de 2 heures ce temps de traitement moyen, sans modifier les processus en place.

Lors de ce premier "tour de roue", les personnes identifiées pour mener l'initiative ETL 2.0 pourront s'approprier la méthodologie du cercle vertueux de l'extrême performance, et ainsi préparer sa pérennisation.

Idéalement, l'initiative doit revenir à une équipe transverse d'expertise aux projets. Régulièrement sollicitée lors des allongements de temps de traitements, en contact avec les services financiers et donc à même d'identifier les applications à fort potentiel en matière de réduction de coûts, ce centre d'expertise appréciera de disposer de leviers simples et industriels pour répondre aux "pains" sur lesquelles elle est régulièrement sollicitée.

Méthodologie ETL 2.0

Correctement déclinée, une initiative ETL 2.0 améliore les niveaux de services des applications et/ou réduit leurs coûts de possessions. Si ses premiers pas seront ciblés sur un domaine applicatif précis, la réelle valeur sera dégagée sur la durée, lorsque les initiatives d'accélération seront menées de manière cyclique. Afin de ne pas mettre en danger les applications et les interfaces inter-applicatives et de produire régulièrement de la valeur mesurable, ces cycles respecteront un mode opératoire cyclique assimilable à une évolution projet.



Ci-dessous un bref descriptif des différentes phases d'un cycle d'accélération

Phase	Descriptif
Surveillance des opérations	- Enquêtes auprès des exploitants pour identifier les processus fortement consommateurs de temps et de ressources - Enquêtes auprès des études pour identifier les projets ayant initié des phases d'optimisation - Enquêtes auprès des utilisateurs pour identifier les services générateurs de mécontentements - Enquêtes auprès des décideurs afin de classer les projets par ordre de criticité stratégique - Enquêtes auprès des financiers afin d'identifier les éléments présentant les coûts de possessions les plus importants

Identification opérations impactantes	- Analyse des enquêtes menées à l'étape précédente en vue d'en tirer un premier ensemble de candidats - Classement des candidats par potentiel d'accélération - Classement des candidats par impact métier, prise en compte de la criticité et du niveau stratégique - Etablissement de la liste explicite des flux considérés dans l'initiative (comité de validation métiers / projets / exploitants)
Initiative d'accélération	- Développements – Migration / développement de tout ou partie des logiques fonctionnelles sur les socles accélérateurs - Tests fonctionnels - livraison de résultats identiques au processus de départ - Tests de performances – Comparaison des processus avant et après accélération - Production - Mise en production des logiques fonctionnelles accélérées
Réduction des temps d'exécutions	Suivi d'exploitation en vue de relever les temps de traitements après accélération et de les comparer à ceux habituellement constatés avant l'initiative
Réduction des ressources utilisées	Suivi d'exploitation en vue de relever les indicateurs de consommation de ressources après accélération et de les comparer à ceux habituellement constatés avant l'initiative

Ainsi menées, ces initiatives cadrées en terme de planning et d'objectifs seront répétables jusqu'à un rythme de une par trimestre, influant ainsi de manière continue, visible, et flexible sur les SLA et les coûts de possessions.

En Conclusion

ETL 2.0 - Pourquoi l'envisager ?

- Parce que les solutions standards du marché ne suivent pas les évolutions de volumes de données actuels
- Parce que les solutions standards du marché utilisées sur de gros volumes de données impliquent des coûts d'acquisition, de possession, et d'évolution inadaptés
- Parce que les approches habituelles de résolution des problèmes de performances impactent

significativement les projets (pilotage, maintenabilité, coûts de possession)
- Parce que les Big Data arrivent

ETL 2.0 – Comment l'implémenter ?

- Ajout dans l'architecture d'une brique accélératrice industrielle transverse, significativement
    - Significativement plus performante que les solutions standards
    - Significativement moins couteuse que les solutions standards
    - Adaptées aux données actuelles et futures
- Déroulement d'initiatives accélératrices vertueuses ciblées
    - Assimilables à des initiatives tactiques
    - Au service d'enjeux stratégiques
    - Transverses
    - Résultats mesurables

ETL 2.0 – Pour quels bénéfices ?

Apporter aux projets des performances améliorées de manière ciblée, industrielle et transverse
- Amélioration tactique des SLA

Réduire de manière significative les ressources allouées aux éléments des infrastructures d'intégration de données
- Réduction stratégique des coûts de possessions

Revenir de comportements accélérateurs ayant entraîné des dérives malicieuses
- Amélioration de l'exploitabilité et de la maintenabilité
- Réduction des besoins en expertise