ETL

Définition

Les outils d'ETL, sigle signifiant Extract Transform Load en anglais (traduit par Extraction, transformation et chargement), est un procédé informatique permettant d'effectuer des synchronisations massives d'informations depuis des sources diverses (bases de données, fichiers) vers des cibles préalablement définies.
Le processus est le suivant : on extrait des données des bases de données de production (l'extraction). Puis, on les transforme pour effectuer des calculs, pour les enrichir avec des données externes (la transformation). Enfin, on charge les données dans les différentes applications décisionnelles (le chargement).

Les outils ETL sont utiles dans le système global décisionnel. Ils sont très importants et pertinents, permettant très souvent d'éviter les échecs, les dépassements de budget d'un projet par exemple.

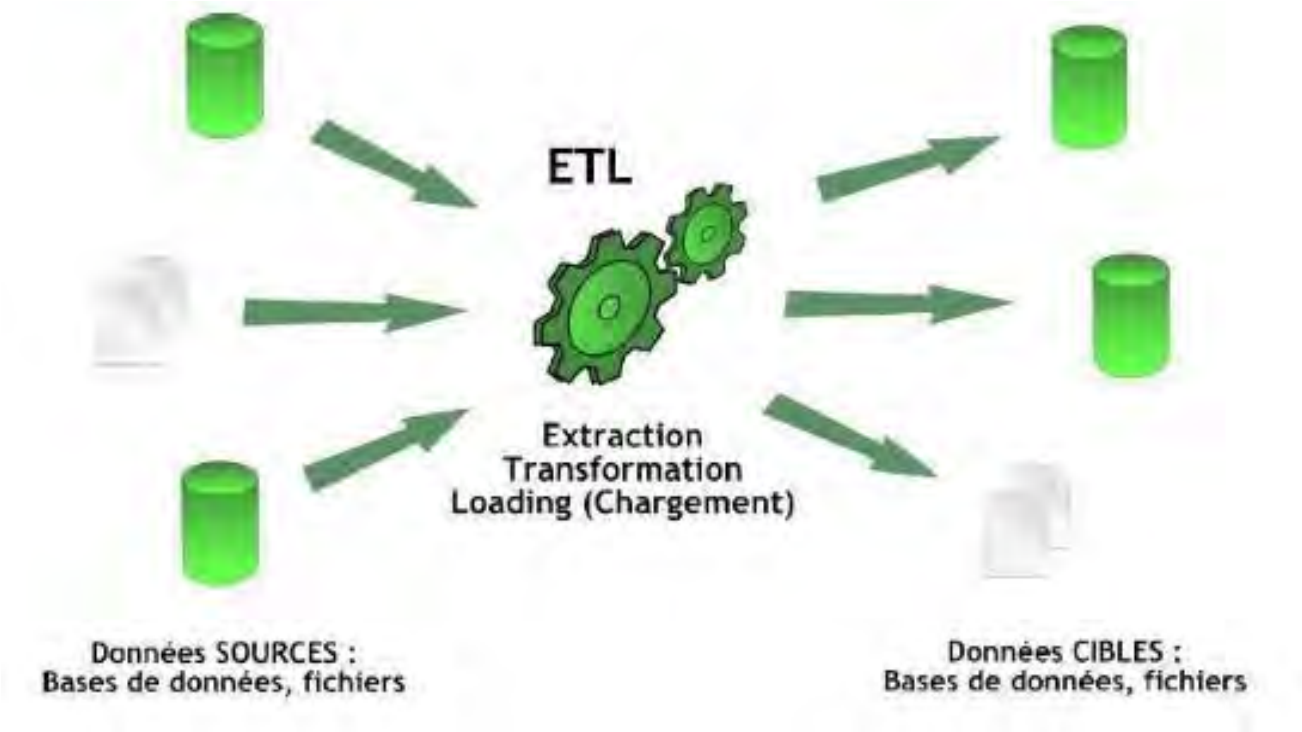

Figure 0-1 : Présentation 1 des ETL

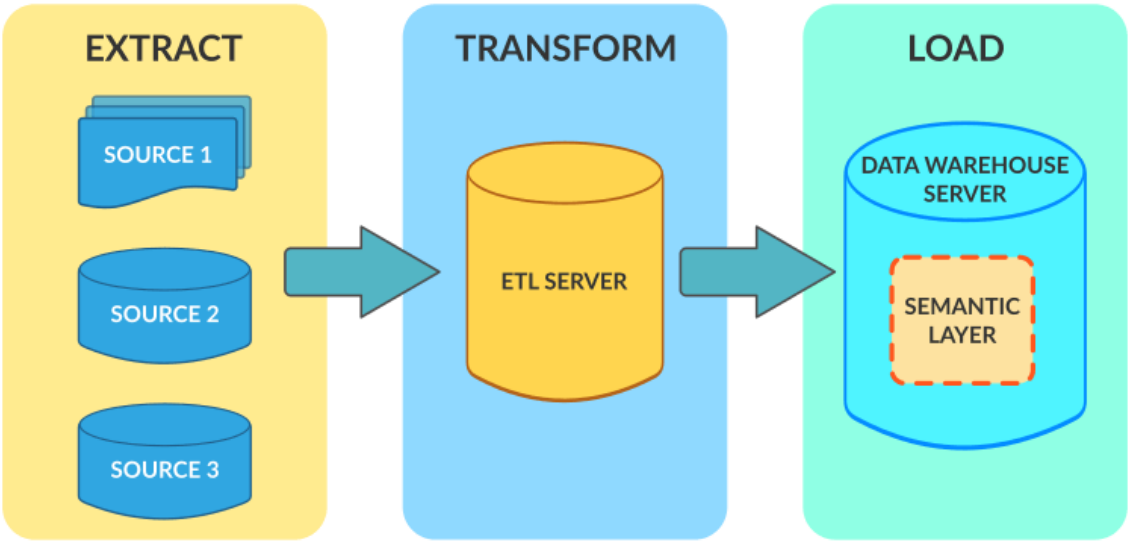

Figure 0-2 : Présentation 2 des ETL

De nombreux systèmes de gestion de bases de données sont supportés nativement en lecture/écriture (Oracle, MS Sql Server, DB2, Postgresql, MySql,...).

De nombreux types de fichiers peuvent également être lus ou écrits: Csv, Excel, Txt, Xml, ...

Notons que la plupart des ETL disposent d'une interface graphique permettant l'élaboration des différents scénarios d'intégration.

Le travail des développeurs en est ainsi grandement facilité, tant au niveau de la conception que de la maintenance des traitements de données.

Les ETL sont communément utilisés dans l'informatique décisionnelle afin de permettre l'alimentation des datawarehouses (entrepôts de données).

Ces derniers servent de supports pour l'analyse des données sous plusieurs formes :

- ✓ Rapports et états,
- ✓ Tableaux de bords (dashboards, balanced scorecard),
- ✓ Indicateurs de performance (« KPIs »),
- ✓ Analyse multi-dimensionnelle (OLAP) ,
- ✓ Analyse exploratoire (Data-Mining).

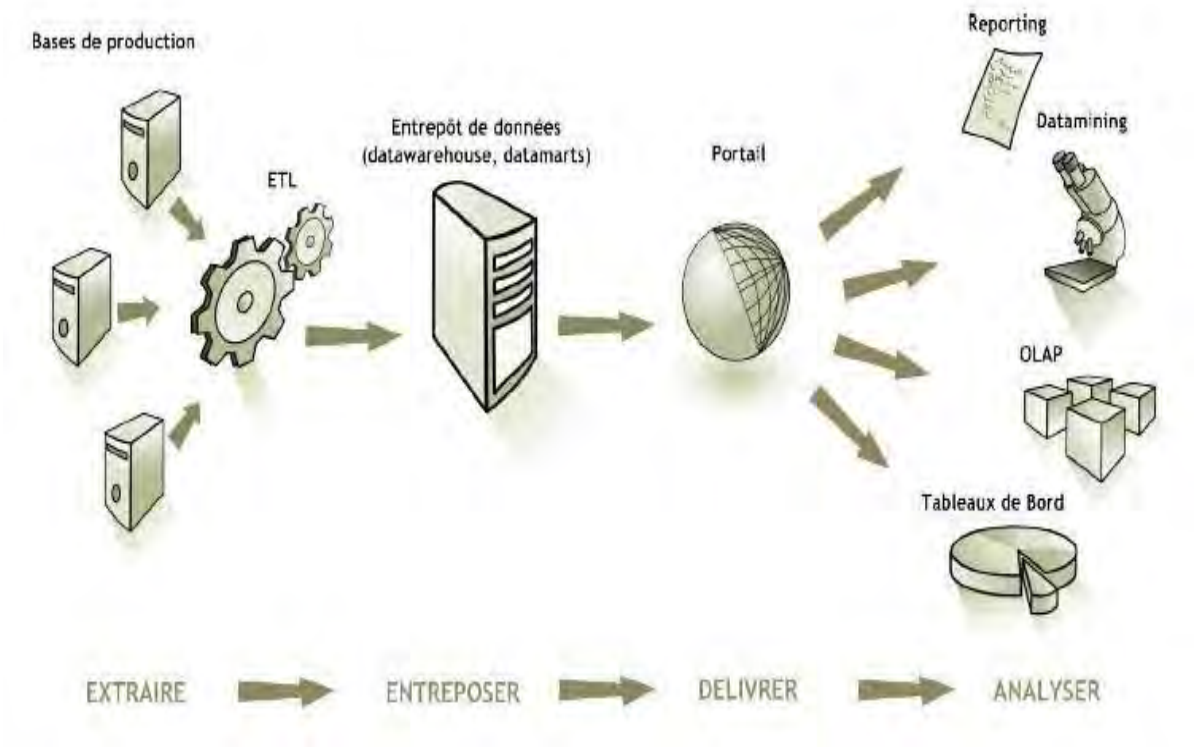

Figure 0-3 : De extraction à l'analyse

Ainsi, les ETL proposent dans la plupart des cas des fonctionnalités propres à l'alimentation des datawarehouses :

- ✓ Création de clés techniques (« Surrogate keys »)
- ✓ Mise à jour de type « dimension Lente » (« Slow Changing Dimension »)
- ✓ Agrégation de données
- ✓ Alimentation de base multidimensionnelle (tables de faits et dimensions)

Les volumes de données traités sont toujours très importants, ainsi les critères de performance sont primordiaux dans le choix d'un ETL.

1. Extract

C'est l'accès à la majorité des systèmes de stockage de données (SGBD, ERP, fichiers à plat...) afin de récupérer les données identifiées et sélectionnées. Prendre en compte les questions de synchronisation et de périodicité des rafraîchissements.

Aujourd'hui, les entreprises disposent de nombreuses sources de données. Les données qu'elles gèrent sont à la fois beaucoup plus nombreuses d'un point de vue volumétrique / quantitatif, et beaucoup plus diverses d'un point de vue qualitatif. Les

entreprises ont potentiellement des dizaines de sources de données et bases de stockage différentes, elles récupèrent des données depuis leur site web, leur site e-commerce, leur ERP, leur système de caisse, leurs outils d'analytics, les réseaux sociaux, les outils de gestion des tickets clients type Zendesk, le CRM, le Marketing Automation, le logiciel d'emailing, les objets connectés, le mobile, etc. On pourrait poursuivre encore assez longtemps cette liste à la Prévert. Le développement des logiciels SaaS, si précieux qu'ils soient, à en partie contribué à multiplier les sources de données. L'augmentation du volume et de la diversité des données est à la fois une aubaine pour les entreprises et un défi. Les outils ETL ont pour vocation d'aider les entreprises à surmonter ce défi. Les données sont éparpillées dans un nombre important de silos. Les logiciels ETL sont des composantes d'une architecture IT visant à remédier au silotage des données.

Le travail premier d'un outil ETL est d'extraire toutes les données en provenance de la myriade d'outils et de bases utilisés par l'entreprise – via des connecteurs et des APIs. Pourquoi faire ? Pour les transformer ! (pourquoi faire ? on y vient rapidement…). Petite remarque : on pourrait dire « collecter » à la place d' »extraire ». Malgré tout, « extraire » convient mieux, car les données que récupère l'outil ETL sont déjà dans l'entreprise, dans le SI de l'entreprise. Il ne s'agit pas de collecter des données nouvelles auprès des utilisateurs, des clients, mais d'extraire des données qui sont déjà dispersées dans le SI de l'entreprise et de les copier sur un serveur virtuel…pour en faire quelque chose : les transformer. Dernière remarque : il est possible de mettre en place des règles (sources à piper, vitesse de rafraîchissement, sources prioritaires…) pour définir les données à extraire. Un outil ETL n'extrait donc pas forcément TOUTES les données, même si il peut le faire.

La plupart des projets d'entreposage de données consistent à consolider les données provenant de systèmes sources hétérogènes, donc différents. Chaque système peut également utiliser une organisation différente des données, comme les règles de nommages, le format des données, etc. En général, l'objectif de la phase d'extraction consiste à convertir les données dans un format unique qui est approprié pour un traitement de transformation. Ce format de stockage est très similaire à la source, on parle de « 1 pour 1 », c'est à dire que l'on souhaite conserver une image des données sources telles qu'elles arrivent. On y applique souvent un système intelligent d'historisation et de journalisation pour savoir quelles données sont arrivées et quand, ce qui permet de reprendre des chargements et gérer les erreurs de transfert de données depuis les systèmes sources.

Transform

L'étape de transformation applique une série de règles appelées règles de gestion, toutes les données ne sont pas utilisables telles quelles. Elles méritent d'être vérifiées, reformatées, nettoyées afin d'éliminer les valeurs aberrantes, celles extérieures à la plage de vraisemblance et les doublons. Puis elles sont consolidées. Il s'agit aussi d'accorder un traitement particulier aux données manquantes. Comment consolider une information si quelques-unes des données la constituant ne sont pas collectées puisqu'elles sont inexistantes dans les bases ?

Les travaux de transformation tels que la standardisation des différents référentiels (unité, échelle) sont d'une ampleur nettement plus conséquente que ne pourrait le laisser supposer ce simple énoncé. Ensuite et seulement on peut procéder aux indispensables agrégations, c'est à dire la fusion de plusieurs données pour obtenir une seule information utilisable (moyenne, somme...).

Ainsi les entreprises ne parviennent plus à avoir une vision globale, unifiée, à 360° de leurs données clients. Trop de données, et surtout trop de sources de données, trop d'outils, trop de bases, trop de formats différents. L'objectif n'est pas de réduire le nombre d'outils qui compose la stack marketing. En tous cas ce n'est pas l'objectif des outils ETL. Les outils ETL, pour le dire très simplement, servent à mettre de l'ordre dans tout ce méli-mélo de données, à structurer les données non structurées, à restructurer des données déjà structurées mais pas comme il faut, à nettoyer les données, à les compiler, les agréger…bref, à transformer, sur un serveur intermédiaire, des données brutes en informations exploitables.

A ce stade, il faut bien comprendre la distinction entre *structurer* et *transformer*. Prenons l'exemple d'une enseigne de grande distribution qui dispose à la fois d'un site e-commerce et de points de vente physiques. L'outil d'ETL utilisé par l'entreprise va extraire, ingérer les données en provenance du site e-commerce et des points de vente physiques. Il s'agit bien plus que de structurer ces données d'ailleurs, certaines le sont déjà. Toutes les données clients ne sont pas des données non-structurées, même si les données non-structurées constituent une part de plus en plus importante des données totales utilisées par l'entreprise. Au-delà du travail de structuration, les logiciels ETL transforment ces données pour que :

- ✓ Elles puissent être exploitées / analysées ensemble : pour que, par exemple, l'enseigne de grande distribution puisse analyser de manière agrégée le comportement de ses clients physiques et celui de ses clients e-commerce.

- ✓ Elles puissent être utilisées par la base de données marketing de l'entreprise. Les outils ETL « préparent » les données. Cette notion de préparation est la notion essentielle. On transforme pour préparer, et non l'inverse. Préparer les données consiste à leur donner un format adapté pour les cas d'usage que l'on souhaite implémenter dans l'outil qui sert de base de données marketing. Les outils ETL rendent les données compatibles *entre elles* et *avec* le référentiel marketing cible. D'un point de vue technique (dans cet article, nous essayons d'être le moins technique possible…), cela consiste à agréger des tables de données entre elles pour créer une « super-table ». C'est comme si, en gros, vous aviez plusieurs tableaux Excel et que vous les agrégiez entre eux pour tout réunir sur un seul tableau et avoir ainsi une vue d'ensemble. C'est le travail que fait l'outil ETL. La forme que prend le « super tableau » et les données qu'il utilise dépendent des cas d'usage voulus par l'entreprise.

Cependant pour certaines sources de données, il faudra une manipulation très légère voire aucune sur les données. Dans d'autres cas, un ou plusieurs des types de transformation suivants peuvent être nécessaires pour répondre aux besoins technico-commerciaux et de la base de données cible:

- ✓ Sélectionner uniquement certaines colonnes à charger.
- ✓ Traduire les valeurs codées, par exmple A2017 pour année 2017 de viendra 2017 ou vice-versa.
- ✓ Joindre les données provenant de sources multiples, par exemple rattacher le nom du client au numéro de sa carte fidélité qui sont dans 2 tables différentes.
- ✓ L'agrégation, avec un cumul du total des ventes pour chaque magasin par exemple.
- ✓ Transposition ou pivot, c'est à dire basculer plusieurs colonnes en plusieurs lignes ou vice versa.
- ✓ Fractionnement d'une colonne en plusieurs colonnes.
- ✓ La suppression des données redondantes, par exemple si l'adresse du client est stocké sur toutes les lignes des ventes, dans la table « Ventes », alors on conservera le numéro de client uniquement et l'adresse du client sera stockée de manière unique dans la table « Clients ».
- ✓ Rechercher et valider les données pertinentes à partir des tables ou des fichiers référentiels pour les dimensions à variation lente.
- ✓ L'application de toute forme de validation de données simple ou complexe.

Si la validation échoue, cela peut entraîner un rejet complet ou partiel des données, et donc certaines ou toutes les données seront remises au prochain chargement, en

fonction de la méthode de conception de l'entrepôt de données et des règles de gestion pour les exceptions. La plupart des transformations ci-dessus peuvent entraîner des exceptions, par exemple, quand une traduction des codes analyse un code inconnu dans les données extraites.

3. Load

Insérer les données dans le Data Warehouse ou le Data Mart.

Elles sont ensuite disponibles pour les différents outils d'analyse et de présentation que sont le Data Mining, l'analyse multimensionnelle OLAP, les analyses géographiques, les requêteurs et autres reportings et bien sûr les tableaux de bord.

Une fois sélectionnées, transformées, formatées, compilées, réconciliées, agrégées, préparées, les données sont envoyées / chargées dans un outil qui sert de base de données marketing. On l'a dit, en général, il s'agit d'un Data Warehouse. D'ailleurs, les deux vont nécessairement de pairs. Il n'est pas possible de monter un Data Warehouse sans utiliser un outil ETL. Car, dans un Data Warehouse, les données sont organisées comme dans un entrepôt. C'est le contraire d'un Data Lake, qui est une sorte d'endroit fourre-tout qui ingère toutes les données sans aucune organisation. Un Data Warehouse stocke des bouteilles d'eau ; un Data Lake ressemble plutôt à un plan d'eau justement. Grâce aux outils ETL, toutes les données stockées dans le Data Warehouse sont, en quelque sorte, disposées sur le même plan, sur le même niveau, en utilisant un langage commun. Il n'y a pas d'étages. Cela facilite la combinaison des données entre elles, l'exécution des requêtes, les travaux d'analyse et permet à l'entreprise de disposer de données de référence, d'une « single source of truth ». Un travail de Business Intelligence permet d'ajouter une couche d'intelligence dans le Data Warehouse et de créer des agrégats permettant de mieux exploiter les données stockées.

II. Evolution

L'ELT existe depuis un certain temps, mais il a connu un regain d'intérêt avec des outils comme Apache Hadoop, un framework de distribution et de traitement des charges de travail volumineuses sur quelques nœuds

seulement ou bien des milliers, pour un traitement parallèle. Les vastes tâches comme la transformation de pétaoctets de données brutes ont été scindées en tâches plus petites, traitées de manière distante, puis renvoyées pour chargement dans la base de données.

Mais les évolutions de la puissance de traitement, en particulier le clustering virtuel, ont permis une augmentation exponentielle de la puissance des ressources de serveur local, réduisant ainsi la nécessité de scinder les tâches. Les tâches Big Data qui étaient habituellement distribuées dans le Cloud, traitées et renvoyées peuvent désormais être traitées en un seul emplacement.

Les plateformes d'intégration de données comprennent d'abord des outils ETL (extract, transform et load) dont la mission est d'automatiser les tâches d'extraction des données multi-source, leur conversion dans des formats adaptés et leur chargement dans des entrepôts de données ou autres bases de données. La première génération d'ETL étaient essentiellement de simples, mais couteux, outils de génération de code aux fonctionnalités limitées. D'ailleurs de nombreuses entreprises estimaient qu'il était bien plus efficace de développer leur propre technologie. La seconde génération, quant à elle, était plus fournie d'un point de vue fonctionnel, mais était centrée sur les traitements par lot, sans trop de performances. Du coup, les DSI pensaient que les ETL ne valaient pas le coup, si les performances n'étaient pas au rendez-vous.

Avec le temps, les outils d'ETL ont évolué et ont élargi leurs champs fonctionnels, en matière de développement, de traitement et de possibilités d'intégration. Pour mieux les positionner comme de véritables plateformes de développement, les fournisseurs d'ETL ont équipé leurs outils de gestion de code, de versioning et génération de documentation et de debugging, par exemple.

Avec l'avancée de l'industrie sur le sujet et sa maturité en matière d'intégration de données, des bonnes pratiques ont pu se dégager, se développer et ont été ajoutées aux outils d'ETL en tant que fonctions pré-intégrées. Ces fonctions portent sur la capture des modifications de données, la gestion de la hiérarchie, la connectivité aux données et la vérification de l'intégrité, par exemple. Les performances se sont aussi améliorées grâce à l'usage de la mémoire et du parallélisme.

Des variantes aux ETL se sont aussi mises en place. On parle d'ELT (pour Extract, Load and Transform). Avec ces outils, il n'est plus nécessaire d'avoir un serveur dédié aux fonctions d'ETL. Ils peuvent être déployés à la source des données ou sur des systèmes cibles, selon leur configuration. L'approche ELT permet donc aux utilisateurs de stocker les données brutes en l'état, de les transformer, dans leur totalité ou partiellement, pour les applications de BI ou analytiques, en fonction des besoins spécifiques.

III. Recommandations

1. Analyses

Pour transformer les données en valeur, il est nécessaire d'investir dans la technologie mais ce n'est pas suffisant. Collecter des données tous azimuts sans une logique préalable, une stratégie spécifique peut s'avérer plus risqué que profitable.

Les problèmes de "silos" et de cloisonnement, les délicates questions de nettoyage et de consolidation, le manque de compétence pour évaluer l'importance de données rebuteront les plus tenaces.
Définir le *"pourquoi"*, pour quels besoins d'analyse, avant le *"Comment"* permettra de répondre aux questions essentielles comme : Quelle données collecter ? Quelles données archiver ? Quelles données rapprocher ? Quelles données sécuriser ? Les questions de sécurité et de confidentialité des données, puis de traçabilité, seront posées au plus tôt du lancement du projet.

2. Les principaux usages des outils etl

Dans la plupart des activités de l'entreprise, les données jouent un rôle essentiel : pour réaliser leur potentiel de valeur, elles doivent être déplacées et préparées pour exploitation, et ces opérations exigent les processus ETL. Exemples de cas d'usage pour les outils ETL :

- ✓ Migrer des données d'une application à une autre
- ✓ Répliquer des données pour la sauvegarde ou l'analyse des redondances
- ✓ Processus opérationnels tels que la migration des données d'un système CRM vers un gisement opérationnel ODS (Operational Data Store) afin d'améliorer ou d'enrichir les données, puis de les replacer dans le CRM.

- ✓ Stocker les données dans un data warehouse avant de les importer, les trier et les transformer dans un environnement de Business intelligence.
- ✓ Migrer des applications on-premises vers des infrastuctures cloud, cloud hybride ou multi-cloud.
- ✓ Synchronisation des systèmes critiques

IV. Avantages et Inconvénients

1. Avantages d'un outil ETL

- ✓ Il dispose de connecteurs base de données, webservices et fichiers plats prêts à l'emploi.
- ✓ Il permet de structurer et rassembler l'ensemble des morceaux de code nécessaires aux transferts et aux transformations des données.
- ✓ Il offre une représentation graphique des flux et opérations.
- ✓ Il permet de traiter rapidement un grand nombre de données.
- ✓ Il facilite la maintenance et l'évolution de l'ETL.
- ✓ Il gère nativement l'encryption des données.
- ✓ Il intègre la gestion des métadonnées, la gestion des erreurs, la gestion des processus et de leur hiérarchie, la gestion de la documentation.
- ✓ Il permet un déploiement facile des flux sur un environnement de production.
- ✓ Il est appréhendable par une personne sans connaissances avancées en développement.
- ✓ Il gère le load balancing entre serveurs.
- ✓ Il est extensible à l'aide de script.

2. Inconvénients d'un outil ETL

- ✓ Il nécessite un ou plusieurs développeurs avec des connaissances sur l'outil d'ETL utilisé ou nécessite un temps d'apprentissage.

- ✓ Il est difficilement intégrable à un outil de gestion de versions.

V. Les différents ETL

1. Choix d'un ETL

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

- ✓ Faut-il créer du code spécifique (procédures SQL, code Java ou autre) ?
- ✓ Faut-il acheter un ETL propriétaire (Informatica, Oracle Warehouse Builder, BO Data Integrator ou autre)?

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

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

Il existe désormais une solution alternative: **Utiliser un ETL Open Source**.

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

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

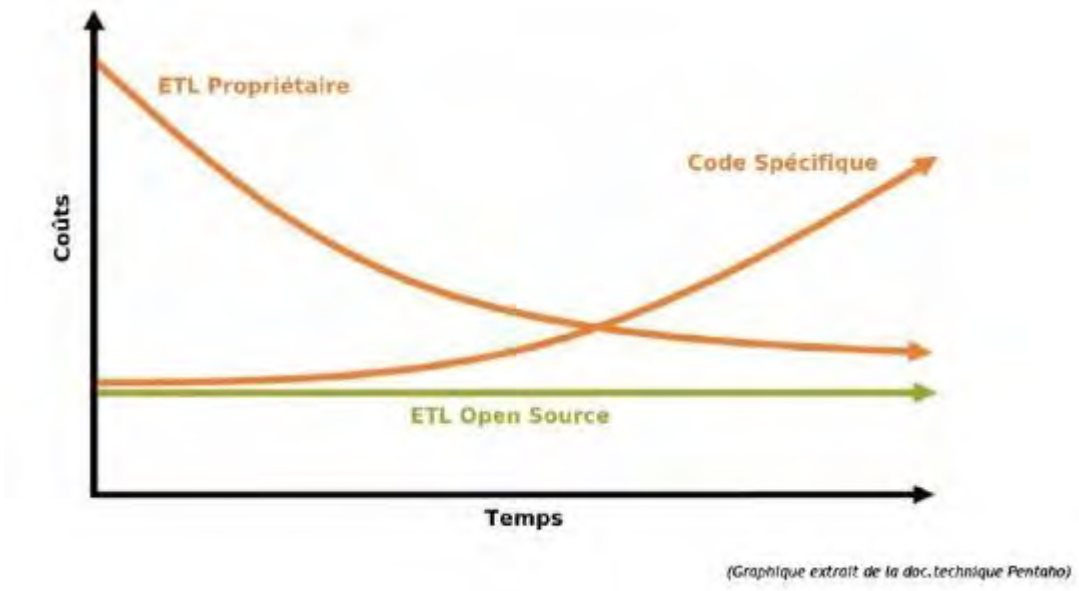

Figure 0-4 : Spécificités des ETL

Figure 0-5 : Les différents ETL

2. Les différents etl

Les termes « Extract, Transform, Load (ETL) » désignent une séquence d'opérations portant sur les données : collecte à partir d'un nombre illimité de sources, structuration, centralisation dans un référentiel unique.

Dans la plupart des entreprises, les données potentiellement utiles sont inaccessibles ; une étude a même révélé que les deux tiers des entreprises retiraient « peu d'avantages concrets » de leurs données, parfois même « aucun avantage ». Les données ont tendance à être enfermées dans des silos cloisonnés, des systèmes legacy ou des applications rarement utilisées. ETL est le processus qui consiste à rendre ces données disponibles en les collectant auprès de sources multiples et en les soumettant à des opérations de nettoyage, de transformation et, au final, d'analytique métier.

Certaines personnes préparent leurs opérations ETL en les codant manuellement en SQL ou Java, mais il existe de nombreux outils pour simplifier ce processus. On examine quelques cas d'usage d'ETL, décrit les avantages des outils ETL par rapport au codage manuel et énumère les qualités à rechercher dans les outils ETL.

a) Propriétaire

- ✓ Anatella
- ✓ Cognos Data Manager
- ✓ DataStudio
- ✓ Dell Boomi
- ✓ Feature Manipulation Engine (FME)
- ✓ Hurence avec un ETL natif Hadoop
- ✓ IBM InfoSphere DataStage
- ✓ Iconics BridgeWorX64
- ✓ Informatica PowerCenter
- ✓ MapReport
- ✓ Microsoft SQL Server Integration Services (SSIS)
- ✓ MyReport Data
- ✓ Oracle Data Integrator de son ancien nom Sunopsis
- ✓ Open executive (éditeur : Cegid)
- ✓ OTIC de son ancien nom OpenText Genio
- ✓ Oxio Data Intelligence solution ETL
- ✓ SAP Data Services qui se nomme aussi Data Integration
- ✓ SAS Data Integration Studio
- ✓ Serenytics

- ✓ Stambia
- ✓ STATISTICA ETL (StatSoft)
- ✓ SynchroDB
- ✓ ect ……..

b) Open Source

- ✓ Talend
- ✓ Vanilla ETL
- ✓ Logstash (souvent utilisé avec Elasticsearch)
- ✓ Pentaho Data Integration
- ✓ CloverETL

VI. Modélisation du processus etl

Un processus ETL regroupe les éléments suivants :

- ✓ Sources de données candidates au processus d'entreposage
- ✓ L'entrepôt de données, destination finale des données à préparer
- ✓ Les tâches d'extraction, de nettoyage, conversion, filtrage, d'agrégation et de chargement

La modélisation devra formaliser tous ces éléments de manière à comprendre le circuit des données depuis les systèmes sources jusqu'à l'entrepôt en passant par les différentes tâches de transformation. Le mappage des données à différents niveaux est un aspect très important pour comprendre le cheminement des données.

1. Modélisation conceptuelle

Le niveau conceptuel consiste à donner une idée très générale sur l'environnement du projet décisionnel à savoir : les besoins des utilisateurs en termes d'analyses (mesures et dimensions), les sources disponibles qui répondent à ces besoins, principales transformations à faire subir aux données sources avant leur chargement dans l'entrepôt. Ce niveau permet aux concepteurs d'aborder les premières réunions pour discuter de la pertinence des données disponibles, leur qualité et sous quels formats existent-elles. Il s'agit aussi de vérifier si tous les besoins peuvent être satisfaits à partir des sources disponibles et quelles sont les transformations nécessaires pour répondre aux exigences de qualité pour l'entrepôt de données.

2. Modélisation logique

Le passage du niveau conceptuel vers le niveau logique consiste à affiner le modèle en intégrant plus de détails dans le modèle logique : séquence d'exécution des activités, différents schémas de données relatifs à une activité (inputs, outputs, paramètres, données rejetées). En dehors du modèle logique, d'autres aspects sont prévus tels que : planification d'exécution du processus, plan de reprise et de restauration en cas de panne, plan d'administration consistant à faire de l'audit, gestion des accès et sécurité, etc.

Il capture les flux de données à partir des sources vers l'entrepôt de données grâce à une suite d'activités synchronisées chargées de préparer les enregistrements de données. Il s'agit de préciser le type des données utilisées lors des traitements, les inputs/outputs et paramètres de chaque activité, la séquence des activités qui montre comment les outputs d'une activité A1 sont utilisés comme inputs dans une activité A2.

3. Modélisation physique

Le niveau physique qui décrit de manière complète le processus ETL doit préciser l'environnement sous lequel s'exécutera le processus : OS, nature des systèmes sources (SGBDs, fichiers plats, Excel, XML, ...), ressources matérielles en termes de serveurs et stations nécessaires ainsi que les profils utilisateurs nécessaires pour faire aboutir le processus de bout en bout jusqu'au rafraichissement de l'entrepôt de données sans échecs.

VII. Comprendre l'utilité de la zone de travail pour les processus etl

L'alimentation d'un entrepôt de données pourrait se résumer en 3 grands points, communément appelés ETL/ETC:

- ✓ **E**xtraction des données depuis les sources dans un environnement de travail,
- ✓ **T**ransformation: Filtrage, nettoyage, manipulation des données pour le chargement de l'entrepôt de données et datamart,
- ✓ **L**oading/**C**hargement: Alimentation de l'entrepôt et mise à disposition des données.

Dans bien des cas, l'utilité de la zone de travail, aussi appelé *staging*, qui intervient entre le "E" et le "T" est trop souvent négligée.
De prime abord, cette étape peut sembler un concept théorique et souvent elle est remise en cause:

- ✓ "Pourquoi utiliser une étape *staging*? il suffit d'alimenter directement l'entrepôt à partir des sources. C'est le travail de l'ETL!"
- ✓ "En se connectant directement sur les sources, on économise une étape, donc du temps d'exécution et de l'espace!"
- ✓ "La *staging*, elle sert à quoi?"

Ces questions fréquentes lors de la phase de conception de l'intégration de données sont monnaie courante et sont la conséquence d'une méconnaissance. Pourtant, la *staging* est une étape nécessaire qui peut bien souvent "sauver" un chargement (journalier) d'un entrepôt de données.

1. Définition du cadre

Son objectif principal est de centraliser des données sources nécessaires en un endroit unique et devient ainsi la source principale de l'entrepôt de données.
Elle garantit que pendant toute la durée d'un chargement, ces données ne sont pas modifiées.

a) L'intégrité

L'avantage d'une image fixe est que le chargement peut s'effectuer en vase clos et s'affranchit complètement du cycle de vie des données dans les sources (données opérationnelles). Ainsi, l'intégrité des sources est garantie jusqu'au prochain cycle de chargement.

À la suite d'un retour ou d'une question utilisateur sur information dans un rapport (par exemple), la *staging* permet d'identifier de manière sûre la donnée source

utilisée pour fournir cette information. Une telle identification permet éventuellement d'apporter une correction de transformation, de tester le résultat et de rejouer l'alimentation sans être à la merci de l'évolution des données dans les systèmes opérationnels.

Un des risques lorsque l'alimentation se fait directement à partir des données opérationnelles est la "perte" d'information: dans encore bien des systèmes opérationnels, la suppression d'une donnée implique une suppression physique de celle-ci. Lors d'un rechargement, cette donnée est perdue et remet en cause de l'intégrité des données de l'entrepôt à cette date et cela peut avoir un impact majeur sur les consommateurs de l'information.

Un autre cas est la mise à jour d'une information dans le système source. Un simple changement de statut peut "falsifier" une information pour les utilisateurs ou fournir des informations incomplètes :

Par exemple, une transaction financière évolue d'un statut "en attente" à "terminée" entre le premier chargement et une reprise: Les informations obligatoires pour une transaction "en attente" est moindre que pour une transaction "terminée" et lors de la reprise, l'alimentation peut s'avérer complexe et coûteuse en temps pour garantir l'intégrité de l'information. Dans bien des cas, le résultat est une transaction où l'information associée à son statut est incomplète.

Bien souvent, les enjeux d'intégrité apparaissent (malheureusement) uniquement en cas de problème alors que le phénomène - le cycle de vie d'une donnée opérationnelle est différent que celui de l'entrepôt - parait évident.

Pour s'en convaincre, aujourd'hui la plupart des entrepôts et autres datamarts contiennent des données journalières alors que les données dans les systèmes sources évoluent dans une même journée. Un cas simple pour illustrer est l'évolution d'un taux de change monétaire dans une journée, au niveau d'un entrepôt, seules les valeurs de clôture sont entreposées.

La différence de cycle de vie d'une donnée justifie pleinement l'utilisation d'une *staging*.

b) La performance

L'extraction des données des systèmes sources peut être synonyme de gros volume et consommateur en ressource machine.

Le chargement d'un entrepôt mixé avec une exploitation directe des systèmes opérationnels est souvent synonyme de *deadlock*.

Pour pallier ce problème, un autre objectif de la *staging* est de stocker ces données sources de manière brute, sans transformations, ni tris, ni exclusions.

Le résultat est un processus d'extraction rapide, car il ne comprend aucune manipulation de la donnée extraite.

Le seul filtre généralement appliqué est celui de la date (date du jour dans le cadre d'un chargement journalier).

L'extraction brute d'information comme les données clients, des portefeuilles, des avoirs minimise le temps d'occupation des systèmes sources sachant que la fenêtre de temps dédiée à l'alimentation d'un entrepôt et l'indisponibilité des systèmes sources est souvent étroite et chaque minute économisée compte.

2. Les points d'attention

- ✓ La *staging* est un environnement "volatile" et les données ne sont conservées que jusqu'au prochain rafraîchissement des données. Cette contrainte implique que dans le cadre d'une alimentation journalière d'un entrepôt, tout problème doit être réglé dans la journée. Toute intervention dépassant le temps alloué entre 2 alimentations peut avoir un impact négatif sur l'activité des usagers exploitants les informations,
- ✓ Les volumes peuvent augmenter de manière exponentielle (nouvelles sources, nouvelles extractions, augmentation des données opérationnelles), il faut veiller à ne pas impacter les temps de chargements et donc le monitorage de la *staging* est une partie intégrante du processus de chargement d'un entrepôt au même titre que le volume des dimensions et des faits,
- ✓ Les tables contiennent rarement des index et donc les performances peuvent être impactées lors de l'exploitation de la *staging* par exemple les informations transactionnelles: trafics réseaux, transactions financières…souvent synonymes de volumes,
- ✓ Malgré la volatilité des données, la *staging* doit être sauvegardée avec la même fréquence que l'entrepôt pour permettre de "rejouer" une journée dans les mêmes conditions que le jour de la première exécution. La sauvegarde implique donc une évaluation des volumes et l'espace requis doit être pris en compte lors de la définition de la stratégie de sauvegarde.

Lors de la mise en place de l'architecture de l'intégration des données, le rôle que doit tenir la *staging* ne doit pas être sous-estimé et celle-ci est une étape importante dans le processus d'alimentation d'un entrepôt.
Dans le cadre des problématiques de chargement, une *staging* correctement organisée est un gage de sécurité et de tranquillité dont il ne faudrait pas se priver.

VIII. Différence entre etl et elt

La différence entre l'ETL et l'ELT réside dans le fait que les données sont transformées en informations décisionnelles et dans la quantité de données conservée dans les entrepôts.

1. Etl

L'ETL (Extract/Transform/Load) est une approche d'intégration qui recueille des informations auprès de sources distantes, les transforme en formats et styles définis, puis les charge dans des bases de données, sources de données ou entrepôts.

Figure 0-6 : Processus ETL

2. Elt

L'ELT (Extract/Load/Tansform) extrait également des données à partir d'une ou plusieurs sources distantes, mais les charge ensuite dans l'entrepôt de données cible sans changement de format. Dans un processus ELT, la transformation des données s'effectue au sein de la base de données cible. L'ELT nécessite moins de sources distantes, uniquement leurs données brutes et non préparées.

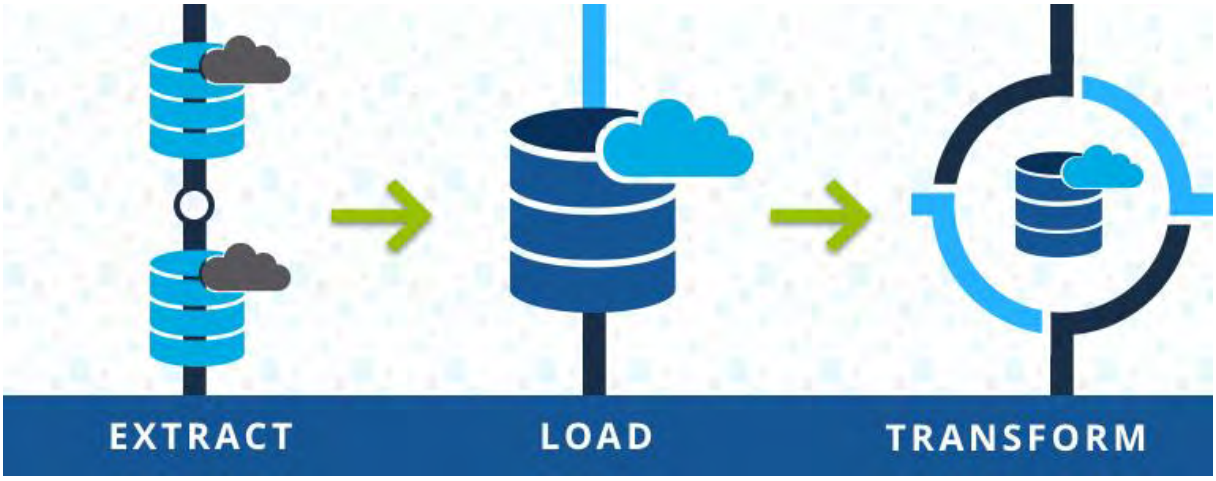

Figure 0-7 : Processus ELT

3. Comment et quand utiliser l'ELT

Contrairement à l'ETL, l'ELT (Extract/Load/Transform) est un processus qui consiste à rassembler les informations provenant d'un nombre illimité de sources, à les charger dans un emplacement en vue de leur traitement, puis à les transformer en données décisionnelles actionnables.

- ✓ **Extraction** – La première étape, l'extraction, fonctionne de la même manière dans les deux approches de gestion des données. Flux bruts de données d'une infrastructure virtuelle, logiciels et applications sont ingérées entièrement ou en fonction de règles prédéfinies.
- ✓ **Chargement** - C'est ici que la branche ELT se sépare de celle de son cousin ETL. Plutôt que de fournir une telle quantité de données brutes et de les charger sur un serveur de traitement temporaire avant transformation, l'ELT livre toutes les données au site dans lequel elles vont ensuite résider. Cela réduit le cycle entre extraction et livraison, mais nécessite un travail bien plus important avant que les données ne deviennent utiles.
- ✓ **Transformation** – La base de données ou l'entrepôt trient et normalisent les données, en en conservant une partie seulement ou la totalité de sorte qu'elles soient accessibles à des fins de reporting personnalisé. La charge de stockage pour une telle quantité de données est plus importante, mais elle offre davantage d'opportunités d'exploration personnalisée pour des données décisionnelles pertinentes en temps quasi réel.

L'approche ELT est-elle le bon choix ? Selon l'architecture réseau existante de l'entreprise, de son budget et de son niveau de maîtrise des technologies

Cloud et Big Data, cela n'est pas toujours vrai. Mais si un ou plusieurs des trois aspects suivants sont stratégiques, la réponse est probablement oui.

a) Lorsque la vitesse d'ingestion prime.

Avec l'ELT, inutile d'attendre que les données soient traitées hors site puis rechargées, (chargement et transformation peuvent s'effectuer en parallèle) ; le processus d'ingestion est donc bien plus rapide, ce qui fournit des informations brutes nettement plus vite qu'avec l'ETL.

b) Lorsque plus de données décisionnelles sont de meilleures données.

La transformation des données en données décisionnelles présente l'avantage de pouvoir révéler et convertir des schémas cachés en informations actionnables. En conservant toutes les données historiques, les entreprises peuvent explorer délais, schémas de vente, tendances saisonnières ou tout autre indicateur émergent qui devient important pour l'entreprise. Étant donné que les données n'ont pas été transformées avant d'être chargées, vous avez accès à toutes les données brutes. Généralement, les data lakes Cloud possèdent un magasin de données brutes, puis un magasin de données affinées (ou transformées). Les data scientists, par exemple, préfèrent accéder aux données brutes, tandis que les utilisateurs métier apprécient les données normalisées à des fins décisionnelles.

.

c) Lorsqu'une évolution est nécessaire.

Lorsque vous utilisez des moteurs de traitement ultra performants comme Hadoop, ou des entrepôts de données Cloud, l'ELT peut s'appuyer sur la puissance de traitement native pour une plus grande évolutivité.

ETL et ELT sont des méthodologies classiquement utilisées pour la production de données décisionnelles à partir de données brutes. Mais, comme pour

tout ce qui concerne la technologie, le Cloud vient modifier la manière dont les entreprises gèrent les problématiques ELT.

4. Résoudre les problématiques ELT fréquentes

Pour une bonne exécution des tâches, chaque entreprise s'entoure des outils et de l'expertise adéquats. Quelle que soit la tâche, les erreurs qui surviennent précocement dans le processus de production sont amplifiées à mesure que le projet prend de l'ampleur et plusieurs pièges courants risquent de nuire aux architectures ELT.

✓ Lacunes de sécurité

En termes de sécurité, il est risqué de déplacer des pétaoctets de données et de les rendre accessibles à toutes les applications et à tous les utilisateurs. Pour une approche de création fiable, la sécurité doit être intégrée à tous les niveaux de l'entreprise, afin de garantir qu'un jeu de données corrompu ou endommagé ne puisse pas infecter les entrepôts de données.

✓ Conformité insuffisante

La mise en place de cadres de conformité tels que HIPAA, PCI et RGPD fait peser sur les entreprises l'obligation de réaliser des audits et de prouver leur respect des normes. Toute approche ELT doit être conçue en gardant à l'esprit les questions de conformité afin d'éviter tout problème avec les règlementations nationales et internationales.

✓ Gonflement des ressources

Si les entrepôts de données présentent des avantages pour explorer les données décisionnelles, ils ont aussi un défaut évident : toutes ces données doivent faire l'objet d'une maintenance. Grâce aux fournisseurs Cloud et à la tarification à l'utilisation, la maîtrise des Big Data est plus abordable que jamais, mais même une tarification différenciée du stockage peut devenir onéreuse sans un plan de gestion qui permet d'éviter la multiplication sans fin des jeux d'informations de travail.

✓ Absence de gouvernance des données

La sécurité des données traversant un processus ELT est essentielle et il en va de même des « 5 W » (Who, What , Where, When, Why) en gouvernance des données :

-Qui contrôle la gestion du master data dans l'entreprise ?
-Quelles sont les données rassemblées/conservées ?
-Quand les présentations et les audits sont-ils exécutés ?
-Où les données sont-elles stockées ?
-Pourquoi les efforts ELT ont-ils un impact positif sur les performances de l'entreprise ?

Anticiper les réponses aux questions clés permet de faire naître des pratiques ELT responsables et préparent les entreprises à des récoltes abondantes d'informations ayant un impact quotidien sur les résultats.

5. Comparaison

- ✓ En utilisant un process ETL, les données qui arrivent dans la base ont déjà une couche d'intelligence. Elles sont exploitables, structurées. En utilisant un process ELT, les données qui arrivent dans le Data Lake n'ont aucune organisation, certaines n'ont aucune structure. Bref, c'est le bazar. Les données sont triées, structurées, normalisées après avoir été chargées.

  Les outils ETL travaillent les données, les structurent, les organisent, en retiennent certaines, en rejettent d'autres, en fonction des besoins de l'entreprise. Toutes les données n'entrent pas dans le DWH. En utilisant le process ELT, à l'inverse, toutes les données atterrissent pêle-même dans la base. Il n'y a aucun traitement, et donc beaucoup plus de données dans le Data Lake. L'entreprise a donc besoin de plus de ressources serveurs. Du point de vue de la Business Intelligence, cela permet de ne passer à côté de rien et de créer des agrégats plus pertinents. Les Data Lakes alimentés suivant un process ELT offrent plus d'opportunités d'exploration.

- ✓ Dans un process ELT, les flux de données sont beaucoup plus rapides : le temps entre le moment où les données sont extraites et celui où elles sont chargées dans la base est plus court. En contrepartie, les données qui arrivent dans la base nécessitent un travail plus long avant d'être exploitables.

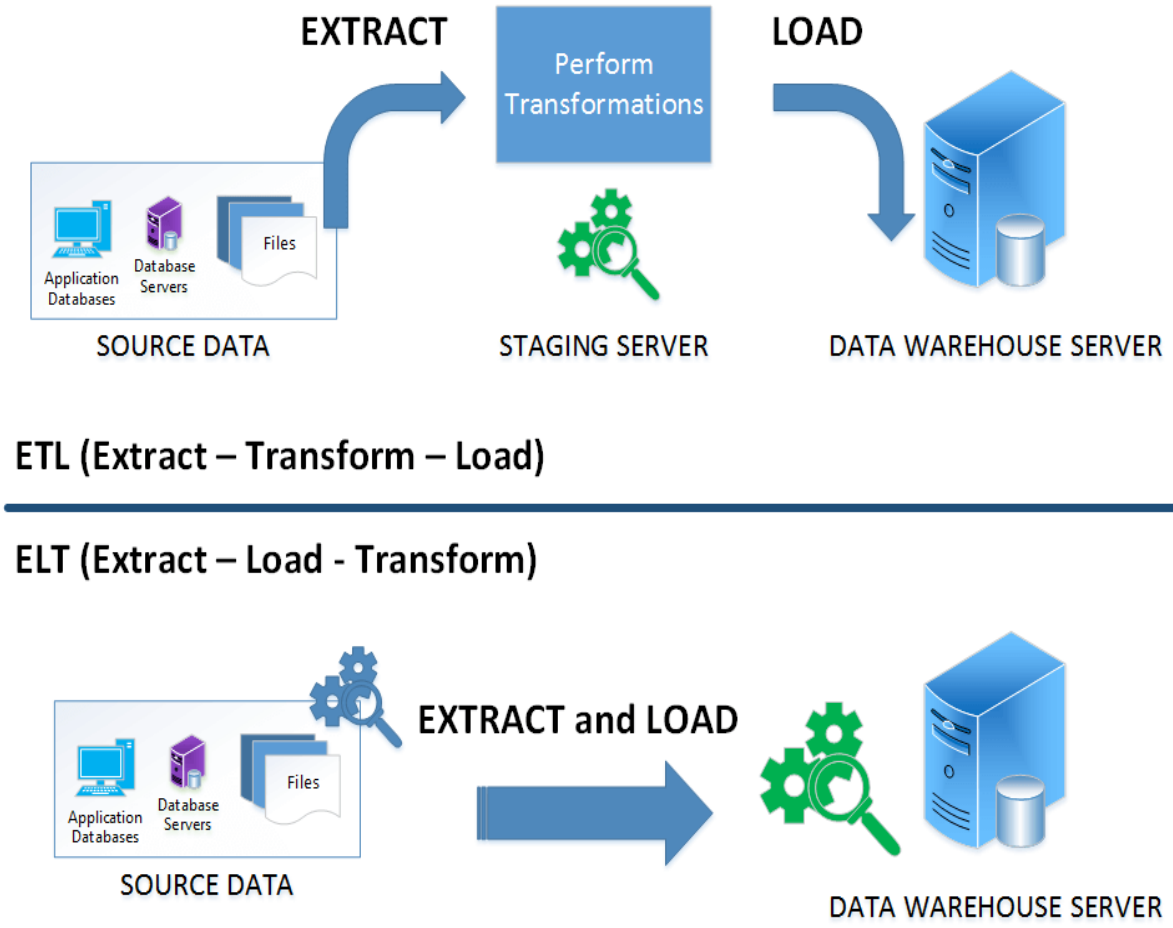

Figure 0-8 : Comparaison etl et elt

Cependant cela ne signifie pas pour autant que l'option ETL soit toujours préférable à l'option ELT. Il y a des cas où cela fait tout à fait sens de privilégier un process ELT. Par exemple, si la vitesse de circulation des données dans le SI est quelque chose de très important pour le succès de votre entreprise, l'ELT peut être envisagé. Comme il n'y a pas de transformation des données, la donnée circule beaucoup plus vite. D'ailleurs, pour gagner du temps, le chargement et la transformation dans la base peuvent être réalisés en parallèle. Deuxième avantage du process ELT : lorsque l'on veut faire de la Data Science, du Machine Learning, identifier des schémas cachés, il est mieux d'utiliser un Data Lake pour ne passer à côté d'aucunes données. Par ailleurs, ce n'est pas parce que l'on juge une donnée inutile à l'instant t qu'elle le sera à l'instant t+1. Dans un Data Lake, on ne perd rien, l'entreprise a accès à toutes les données brutes historiques. Par contre, il est plus difficile de se conformer aux réglementations relatives à la protection des données personnelles (avec ce type d'architecture. De fait, mettre en conformité ces réglementations une architecture Data Lake / ELT est

complexe…et coûteux. Aussi, la sécurité et la maintenance des données sont des sujets complexes dans l'ELT…Pour ces raisons, dans la plupart des cas, l'architecture ETL + Data Warehouse est plus pertinente.