

ETL ou ELT ?

Choisissez le processus
le mieux adapté à vos
objectifs.

L'évolution des livraisons de données

Que la livraison de données n'ait aucun secret pour vous ou que vous commenciez à peine à définir votre stratégie, vous partagez le même objectif que vos pairs : diffuser des données dans toute votre organisation de manière efficace, fiable et reproductible.

Durant de nombreuses années, les opérations ont suivi un ordre bien précis : extraire, transformer et charger – ou ETL. En raison de la quantité de données et du nombre d'outils disponibles, il était tout à fait logique de travailler avec ce modèle. D'ailleurs, actuellement, il continue de bien fonctionner avec certains cas d'usage.

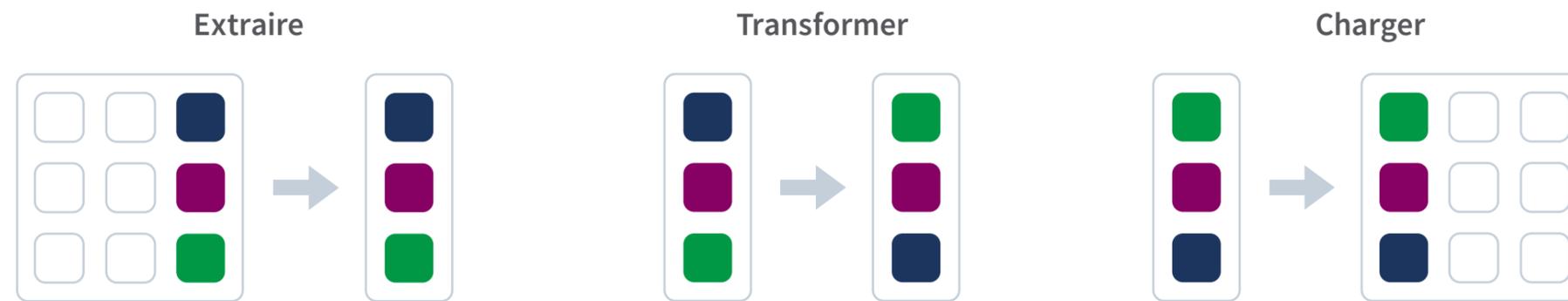
Toutefois, avec la multiplication des données, l'accélération de la vitesse, la diversité des activités et l'arrivée de nouveaux outils d'intégration, les utilisateurs se tournent de plus en plus vers une nouvelle alternative : extraire, charger et transformer – ou ELT. Le modèle ELT offre de nouveaux avantages par rapport à l'ETL, en particulier lorsque la vitesse est une priorité.

Comment savoir quel modèle utiliser ? Cet e-book vous donne la réponse.



Les avantages de l'ETL

L'ETL est l'ensemble des processus impliqués dans l'extraction des données d'un système, leur transformation et leur chargement dans un référentiel cible. Il s'agit d'un flux de travail traditionnel qui permet de nettoyer, enrichir et transformer les données avant de les intégrer et de les exploiter dans le cadre de l'analytique, de la Business Intelligence et de la data science.



Avec l'ETL, la transformation a lieu dans un zone de synchronisation, avant le chargement dans le data warehouse. Le jeu de données complet doit être transformé avant son chargement, mais une fois ce dernier effectué, une analyse systématique et précise peut immédiatement avoir lieu.

Les principaux avantages de l'ETL sont :

- **La maturité :** l'ETL est une fonction bien comprise et bien prise en charge.
- **L'accélération de l'analyse pour un cas d'usage prédéfini unique :** le jeu de données ayant été structuré et transformé au préalable, l'analyse a lieu immédiatement.
- **La transparence :** dans certains cas, la traçabilité des données est simplifiée, car la transformation des données est comprise dans le processus, au lieu de survenir ultérieurement.

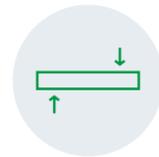


Les limites du modèle ETL

Si le modèle ETL reste une solide option pour certains cas d'usage, il présente des inconvénients, en particulier en raison de l'augmentation du volume de données et du rythme des activités. L'ETL est :



Chronophage : surtout si vos développeurs doivent écrire le code. Le processus de transformation doit suivre un ensemble d'exigences dont la formulation peut prendre des semaines, voire des mois. Après la transformation, le chargement des données peut égaler cette durée. Il existe des outils spécialement conçus pour le modèle ETL qui réduisent ou éliminent complètement le codage, afin d'accélérer le processus. Toutefois, ils ne sont pas d'un grand secours pour optimiser un pipeline opérationnel d'intégration de données.



Rigide : l'ETL ne prend pas directement en charge les flux de travail. Pour que ce modèle fonctionne, tous les paramètres doivent être définis, puis mis en œuvre. Si un nouveau cas d'usage se présente, il faut reprendre tout le processus ETL à zéro.



Fragile : modifier une étape d'un flux de travail ETL peut invalider d'autres flux de travail car les suivants dépendent souvent des précédents. Autrement dit : Si le premier flux de travail doit s'exécuter entièrement avant que le deuxième ne puisse démarrer, en cas d'échec du premier, le deuxième n'a jamais lieu.



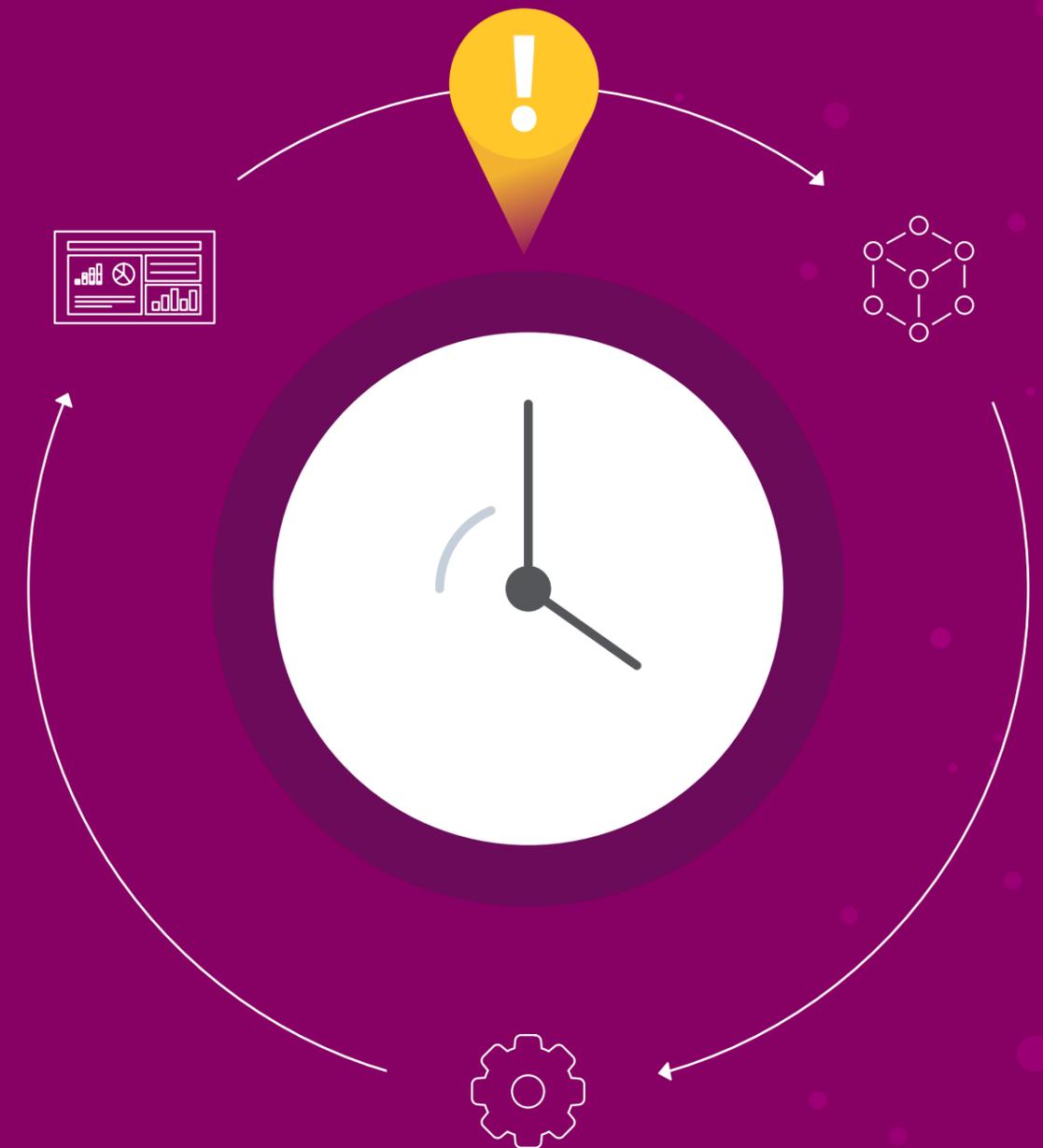
Opaque : le processus de transformation n'est pas transparent puisque seuls les experts qui transforment les données y ont accès.

La question du traitement par batch

L'approche ETL pose un autre problème : les limitations en matière de traitement par lots. Généralement, les processus ETL fonctionnent selon un planning prédéfini : chaque minute, heure, jour ou semaine en fonction du cas d'usage. Ils peuvent aussi s'exécuter en réaction à un élément ou événement déclencheur, bien que ce soit plus rare.

Un processus ETL planifié fonctionne en mode de traitement par batch. Sa fréquence tend alors à dépendre de contraintes telles que la rapidité de la mise à disposition des données requises ; la durée d'extraction, de la transformation et du chargement des données ; et le volume de transactions entre deux lots. Globalement, c'est un processus qui fonctionne bien. Néanmoins, lorsque la quantité de données atteint un certain niveau ou lorsque la durée de l'ETL dépasse la fréquence prévue, tout s'effondre.

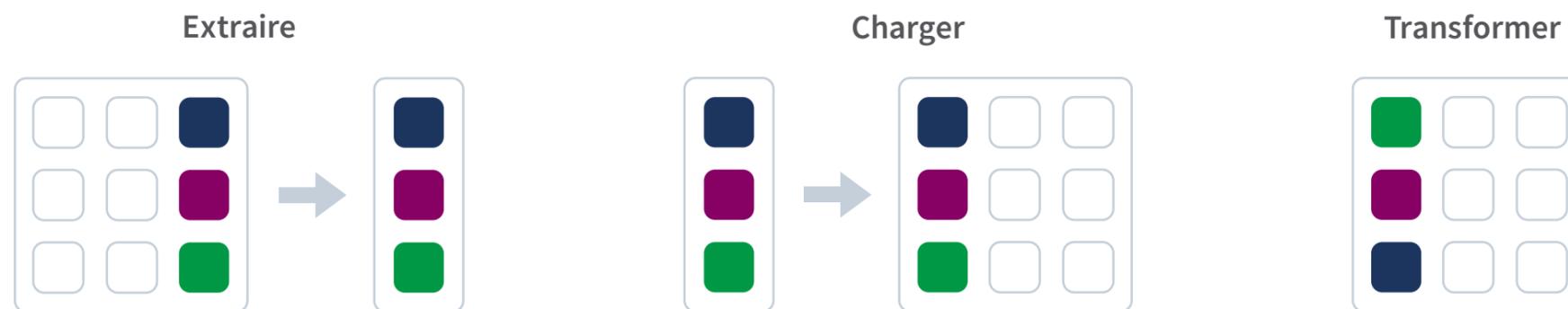
Par exemple : une banque peut avoir besoin de mettre à jour son data warehouse avec un million de transactions toutes les 10 minutes. Mais que faire si extraire, transformer et charger le lot prend 15 minutes ? Étendre la fréquence à 20 minutes ne résout pas le problème. En effet, au bout de 20 minutes, la quantité de données a augmenté de deux millions de lignes, ce qui demande encore plus de temps pour la traiter. Il est évident qu'une autre approche est nécessaire dans de tels scénarios.



L'arrivée de l'ELT

Les organisations actuelles doivent traiter d'énormes quantités de données qui affluent dans différents types de formats. Pour rester compétitives à l'ère de l'économie en temps réel, elles doivent agir aussi vite que possible. C'est exactement le type de scénarios qui met l'ETL en défaut et c'est pourquoi l'ELT suscite un engouement croissant.

Le modèle ELT consiste à extraire, charger et transformer. Il s'agit d'un ensemble de processus d'intégration de données qui permet d'extraire les données d'un système, de les charger dans un référentiel cible, puis de les transformer afin de les exploiter en aval dans le cadre de la Business Intelligence (BI) et de l'analyse du Big Data.



Au cours du processus ELT, la transformation des données a lieu selon les besoins, au sein du système cible. Tout d'abord, un outil d'extraction collecte des données d'une ou de plusieurs sources. Parfois, elles sont stockées temporairement dans une zone de synchronisation au sein d'une base de données. Ensuite, les données sont placées dans le système cible. Il s'agit généralement d'un data warehouse dans le cloud, où les données sont prêtes à être analysées par des outils de BI ou analytiques. Les données sont transformées selon les besoins, dans le système cible : par exemple, selon la structure ou le format d'un jeu de données compatible avec le système cible.

Les avantages de l'ELT

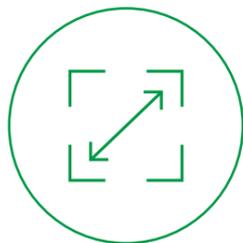
L'ELT améliore l'efficacité du processus de gestion du Big Data et des data warehouses dans le cloud.

Ses principaux avantages sont :



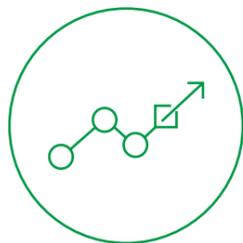
Encore et toujours la vitesse :

le processus ELT implique généralement de charger vos données dans un référentiel dans le cloud. Par ailleurs, les services cloud actuels vous permettent de transformer et traiter rapidement de vastes jeux de données.



La flexibilité :

le processus ELT vous permet de répliquer toutes vos données brutes à la demande, en fonction des besoins de votre analytique, de votre BI ou des autres systèmes.



L'évolutivité :

les data warehouses, data lakes et data lakehouses modernes dans le cloud vous procurent l'évolutivité nécessaire pour exploiter toutes vos données brutes.



La transparence :

les utilisateurs de données peuvent voir quelles données sont disponibles dès leur arrivée ou presque.



Le coût réduit :

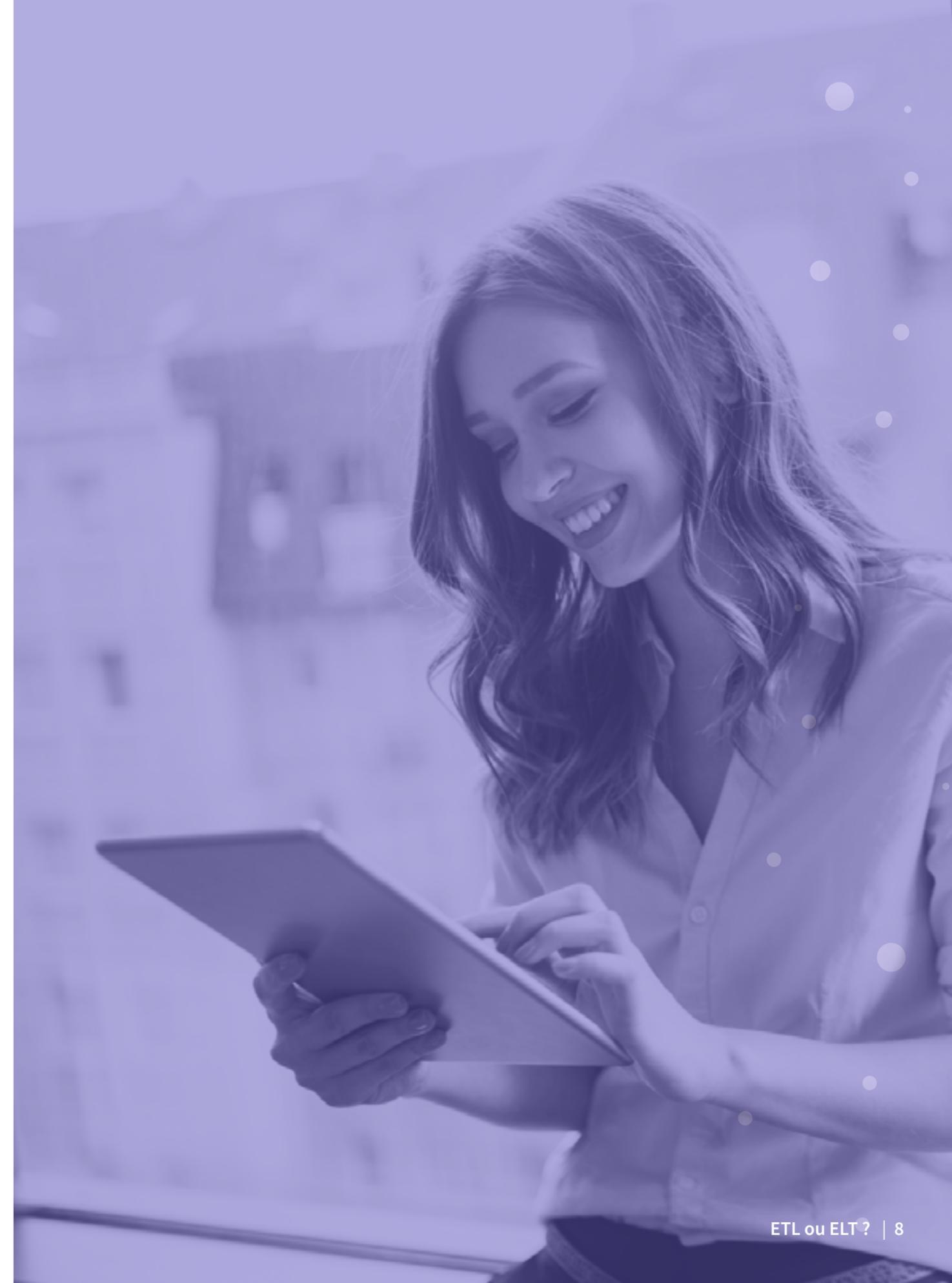
les flux de travail ETL nécessitent un investissement pour du matériel sur site. Avec l'ELT dans le cloud, au contraire, vous payez seulement en fonction de votre utilisation, ce qui réduit généralement le coût total.

ETL + ELT : pourquoi pas les deux ?

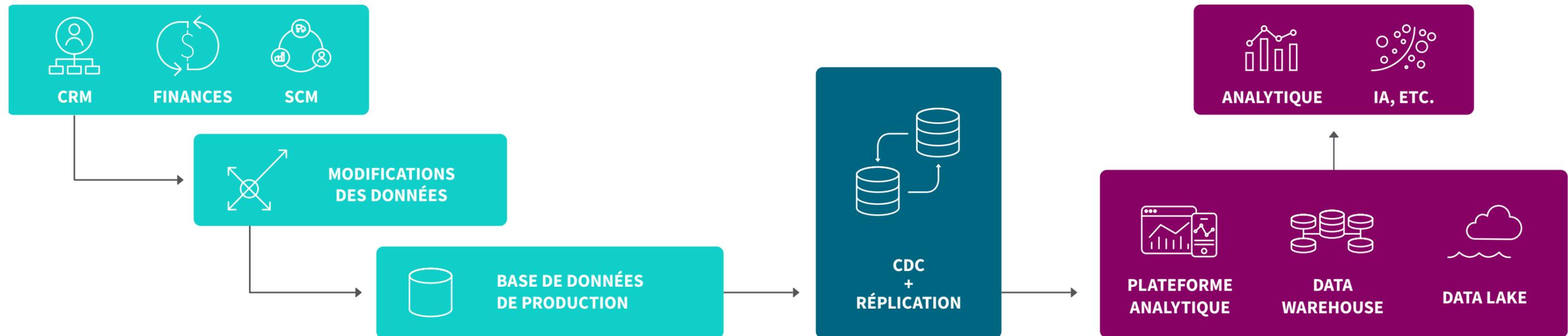
De manière générale, le processus ETL est efficace pour les petits jeux de données qui nécessitent des transformations complexes et ne sont pas requis rapidement. L'ELT est quant à lui plus approprié aux grands jeux de données structurés ou non structurés, lorsque la contrainte temporelle est importante.

De nombreuses organisations utilisent les deux pour couvrir l'éventail de leurs besoins en matière de livraison de données, en passant d'un modèle à l'autre ou en combinant les deux selon les exigences métier. Par exemple, il est possible d'utiliser l'ETL pour des exigences existantes de reporting et l'ELT pour des pipelines de données en temps réel basés sur la technologie Change Data Capture (voir page 9) alimentant de nouveaux processus analytiques, tels que le reporting de fraude et le machine learning.

Toutefois, ce n'est pas sans raison que le marché adopte l'ELT comme nouvelle référence. La quantité, la vitesse et la variété des données ont tellement augmenté ces dernières années que l'ELT a souvent remplacé l'ETL comme modèle de facto pour déplacer les données. C'est en particulier le cas dans des scénarios tels que la migration de données vers le cloud, l'ingestion de données dans les data warehouse et data lakes, ainsi que le MLOps (livraison et automatisation en continu de pipelines de données utilisant le machine learning).



Améliorer l'ELT avec la technologie Change Data Capture



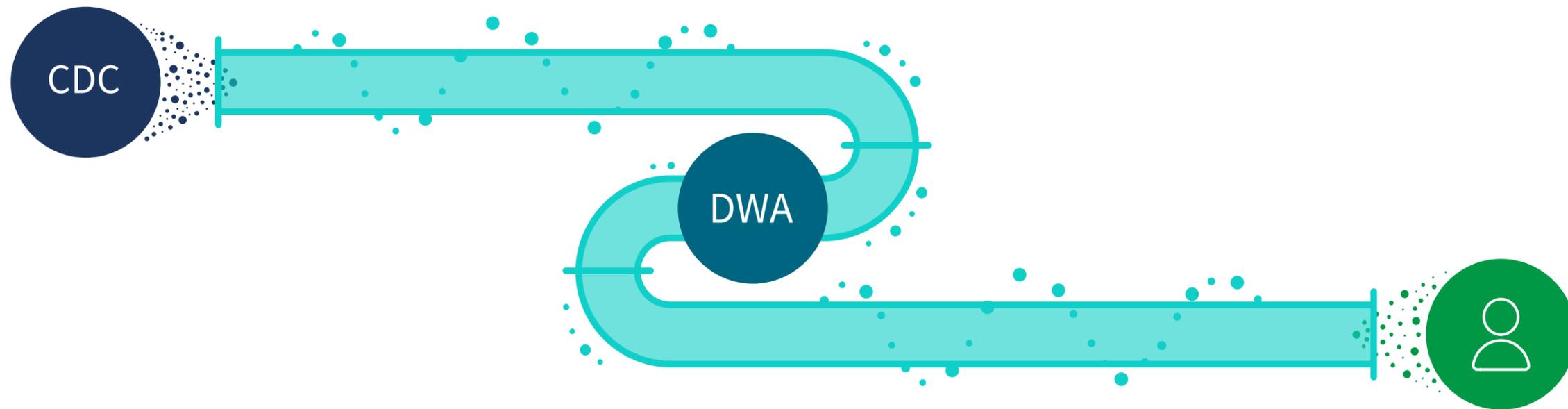
Vous l'avez compris, l'approche ELT se distingue par sa vitesse. Revenons à notre exemple de la page 5 avec une banque et remplaçons l'ETL par l'ELT. Quand extraire, transporter et charger les données prend 15 minutes, l'ELT permet d'avoir les données dans le système bien plus rapidement, parfois en quelques secondes. La banque peut alors choisir de planifier un processus de transformation en masse (idéalement quand il y a plus de ressources disponibles) ou de reporter la transformation après l'utilisation des données.

Il est même possible de faire encore mieux en utilisant la technologie Change Data Capture (CDC), qui permet de répliquer instantanément les modifications apportées aux données. La CDC ne repose pas sur des lots à la fréquence planifiée, comme l'ETL. Elle est activée à chaque fois qu'une modification est apportée à la source de données. Dans notre exemple, les processus EL (sans le « T ») s'exécuteraient à chaque transaction, ne nécessitant la transmission sur le réseau que d'une quantité minimale de données. Sans avoir à attendre le traitement d'un million de lignes. Les processus d'extraction et de chargement auraient véritablement lieu en temps réel.

Que peut apporter la technologie CDC à votre entreprise ?

Découvrez une vue d'ensemble sur notre [page CDC](#).

Élaborer un pipeline de données en temps réel avec l'automatisation du data warehouse



Tout comme la CDC automatise les processus d'extraction et de chargement du pipeline de données, l'automatisation du data warehouse (DWA) prend en charge la transformation. La DWA élimine le codage manuel, propice aux erreurs, et génère automatiquement les commandes, les structures du data warehouse et la documentation nécessaires à la conception, à l'élaboration et à la maintenance d'un programme de data warehouse. Vous gagnez du temps, réduisez vos coûts et limitez les risques liés au projet.

Si l'intégration entre vos solutions de Change Data Capture et d'automatisation du data warehouse est fluide, vous obtenez un ELT en temps réel. En d'autres termes, vous créez un pipeline de données en temps réel. Vous atteignez ainsi l'Intelligence Active, un état de conscience continue qui permet à chaque membre de votre organisation d'agir instantanément en fonction des données, au rythme de leurs changements. Cet objectif, qui durant des années a semblé impossible à réaliser, a toujours été la mission de la Business Intelligence. Aujourd'hui plus que jamais, l'Intelligence Active est absolument essentielle pour obtenir un avantage concurrentiel sur le marché actuel. Heureusement, nous avons désormais la technologie pour la mettre en place.

Comment l'automatisation du data warehouse révolutionne-t-elle la livraison de données ?

Découvrez-en plus sur notre [page DWA](#).

CAS CLIENT SUR L'ELT

Jaguar Land Rover fournit l'ingestion de données-as-a-service avec Qlik®.

LE DÉFI

Sur un marché international complexe, soumis à des réglementations diverses et parfois à des perturbations de la chaîne logistique, Jaguar Land Rover a besoin de disposer de données en temps réel pour prendre des décisions immédiates et avisées. Cependant, les données pertinentes sont verrouillées dans plusieurs systèmes sources essentiels, à la fois sur site et dans le cloud.

LA SOLUTION

L'équipe de cette entreprise dédiée au traitement des données a décidé d'adopter un data lake et d'utiliser la technologie de Change Data Capture/streaming des données pour le remplir. Après l'évaluation de plusieurs solutions, Qlik Replicate™ s'est révélée être la plus efficace (et en fait la seule) pour le transfert de données vers des couches de stockage très hétérogènes dans le lac.

LES RÉSULTATS

L'ingestion-as-a-service est désormais une fonctionnalité utilisable dans toute l'entreprise. Une réplication ne prend que 15 minutes. Les données SAP sont désormais disponibles sur les sites Web externes de JLR pour différents pays. Elles permettent de répondre rapidement aux exigences réglementaires des marchés mondiaux.

[Lire le cas complet.](#)



« Nos données doivent être disponibles pour que nous puissions relever rapidement les défis dès qu'ils se présentent. Qlik était la seule solution capable de fournir une prise en charge aussi vaste sur tous les principaux systèmes de l'environnement de JLR. »

Michael Cockbill
Technical Product Manager

Vitesse, volume et avantage concurrentiel : passez à l'ELT !

À mesure que l'économie se transforme, que les données évoluent et que de nouveaux outils font leur apparition sur le marché, les stratégies data doivent elles aussi s'adapter.

Bien que l'ETL reste efficace pour certains cas d'usage et qu'il soit utile de le conserver dans vos processus, l'ELT offre de nouveaux avantages clés, notamment pour intégrer d'immenses quantités de données rapidement. En effet, le modèle ELT vous permet de créer un pipeline pour agir en fonction des données en temps réel, en plus de constituer un facteur de taille pour accroître la compétitivité de votre organisation.

Vous voulez tester le déplacement de données en temps réel avec une solution ELT primée ? Essayez Qlik Replicate, notre solution de Change Data Capture.

[Essai gratuit](#)

Qlik s'est donnée pour objectif la création d'un monde « data literate », où chacun peut exploiter la data et l'analytique pour résoudre les problèmes les plus complexes. La plateforme d'Intelligence Active dans le cloud de Qlik offre une intégration des données en temps réel de bout en bout et des solutions d'analytique dans le cloud, afin de rapprocher données, informations et actions. En transformant les données en Intelligence Active, les entreprises peuvent s'orienter vers de meilleures décisions, améliorer leur chiffre d'affaires et leur rentabilité, et optimiser la relation client. Qlik exerce ses activités dans plus de 100 pays et offre ses services à plus de 38 000 clients actifs à travers le monde.