Fiche

INTRODUCTION AU DATA PROCESSING

Data lake, ETL, Data warehouse, Batch processing, Data engineer… tous ces termes sont pour vous du charabia ? Pas de panique ! Cette fiche vous propose de découvrir tout ce jargon qui appartient en réalité au **Data processing**, le processus qui **traite vos données** pour les **transformer en information utile** et en **valeur ajoutée** pour votre entreprise.

Découvrez également ce qu'est le **Data engineering**, la discipline en lien, entre autres, avec la **qualité de vos données** ainsi que le **processus ETL**, détaillé étape par étape, et ses spécificités par rapport au **Big Data**. Ensuite, nous examinerons les particularités de deux modes de stockage de données bien connus que sont les **Data lakes** et les **Data warehouses**.

Vous verrez qu'il n'y a pas de solution unique qui soit meilleure que les autres, cela dépend de votre entreprise et surtout de vos besoins.

hub créatif charleroi métropole

http://www.hub-charleroi.be

info@hub-charleroi.be

Partenaires

ULB Université Libre de Bruxelles — IGRETEC — CIE Charleroi Entreprendre — Ocetic — MWS enmieux.be

FEDER — Union européenne — Wallonie

Le fonds européen de développement régional et la Wallonie investissent dans votre avenir

Fiche Introduction au Data processing

Pour quoi faire?

A l'heure actuelle, notre société est bouleversée par ce qu'on appelle la *transformation numérique* : les entreprises intègrent de plus en plus de technologies numériques au sein de leurs activités afin d'accroître leur productivité, leur croissance, leurs innovations,... et donc leur compétitivité.

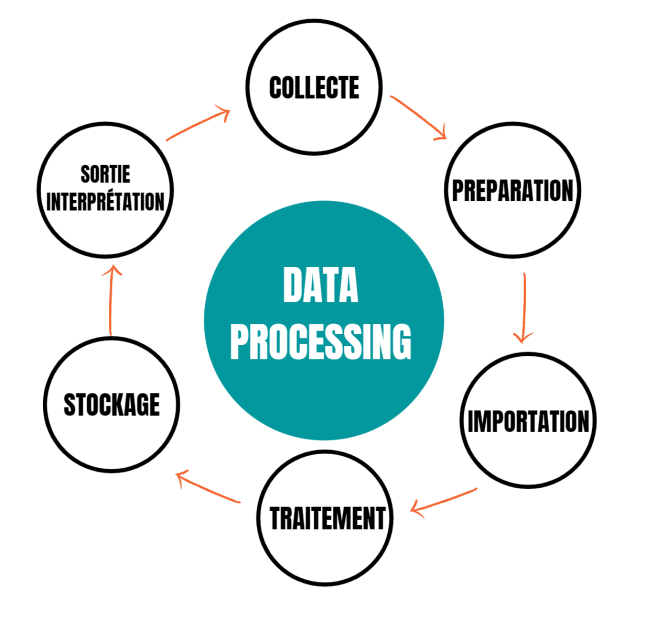
Les entreprises ont maintenant accès à une multitude de données provenant de sources diverses et variées (vidéos, réseaux sociaux, sites consultés, formulaires complétés,...). Ces données doivent être correctement interprétées pour fournir des informations utiles à l'entreprise, c'est-à-dire des informations qui permettent d'analyser la situation passée et présente afin de prendre des décisions pertinentes concernant le futur (résoudre des problèmes, proposer de nouveaux produits,...). Un des grands défis du Big Data est donc de savoir que faire dire aux données et dans quel but ? Comment les interpréter correctement afin de les comprendre, de les analyser, de leur ajouter une valeur et les utiliser pour améliorer l'expérience utilisateur ?

Le Data processing

Le *data processing* est ce qu'on appelle le traitement des données au sens large, c'est le processus qui va transformer vos données brutes en informations exploitables. Le traitement des données se compose des étapes suivantes :

1. La collecte des données
2. La préparation des données (*pre-processing*) : on améliore leur qualité.
3. L'importation: Les données propres sont ensuite saisies dans leur destination.
4. Le traitement des données (*processing*) : les données sont transformées, par exemple à l'aide d'algorithmes d'apprentissage automatique.
5. La sortie et l'interprétation : il s'agit du résultat. Les données sont converties, lisibles et exploitables car elles fournissent de l'information utile. Elles sont présentées sous forme de graphiques, images, vidéos,...
6. Le stockage : les données transformées sont stockées pour être utilisées ultérieurement.

Fiche Introduction au Data processing



Ces étapes forment un cycle qui va vous permettre d'extraire de la valeur de vos données.

Le Data engineering

Bien que certains considèrent que les origines du *Data Engineering* (ou l'*Ingénierie des données* en français) remontent à 1980, voire à 1950, c'est à partir des années 2000 et l'avènement du Big Data que cette discipline devient vraiment nécessaire, et à partir de 2010 que ce terme est popularisé.

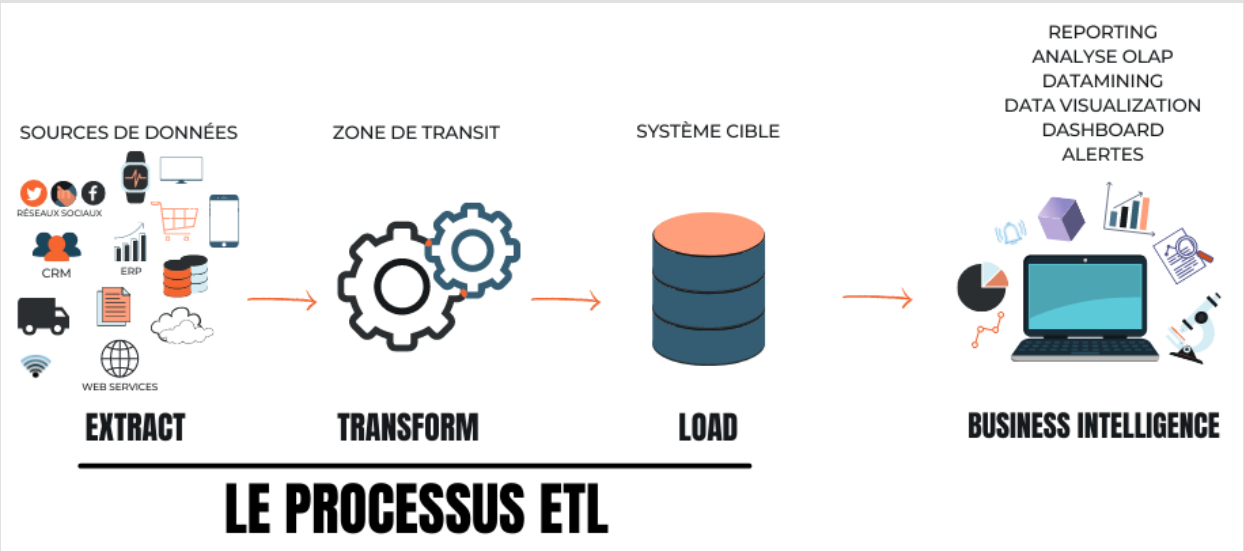
Souvent confondue avec la *Science des don*nées (*Data Science* en anglais), l'ingénierie des données vise à mettre en place les outils et infrastructures nécessaires et adéquats pour l'analyse, la préparation et le traitement des données volumineuses afin d'en garantir leur pertinence, leur qualité et d'éliminer celles qui sont inutiles. L'*ingénieur des données* (*Data Engineer*) fournit les données traitées et prêtes à l'usage aux *Data Scientists* (*experts en Science des données*) qui effectueront de l'*analyse prédictive*, du *Machine Learning* ou du *Data Mining* à partir de celle-ci.

Le *data engineering* répond aux 5V caractéristiques du Big data : afin d'obtenir davantage de connaissances sur ses clients, ses ventes, ses stratégies marketing, ses besoins,... et donc d'obtenir un avantage concurrentiel sur le marché (*véracité* et *valeur*), une entreprise doit ingérer énormément de données (*volume*) provenant de sources diverses (*variété*) et les traiter rapidement (*vitesse*).

Fiche Introduction au Data processing

Le processus ETL

Apparu dans les années 1970, un *ETL* (*Extraction-Transform-Load*) est le processus utilisé par les *ingénieurs de données* (*Data Engineer*) pour transformer vos multiples données brutes en informations commerciales exploitables. L'ETL facilite la migration de gros volumes de données provenant de sources multiples vers un emplacement centralisé afin d'en obtenir une vue globale et unifiée, c'est de l'*intégration de données* (*Data integration*). Cela se fait en 3 phases : l'extraction, la transformation et le chargement.



Fiche Introduction au Data processing

1 Extraction (*extract*)

Aujourd'hui, il y a une multiplication des flux de données et de leur quantité. Davantage complexes, les données proviennent de sources multiples (de votre smartphone, de votre montre connectée, des systèmes d'entreprise, des API, de n'importe quel capteur, d'outils de marketing, de bases de données de transactions, de *data lake* et de *data warehouse*,...) et peuvent donc avoir n'importe quelle structure et format. Il existe 3 types de structure de données :

- Les données *non structurées* : elles ne sont absolument pas organisées et sont sous forme brute absolue. Ce sont par exemple des e-mails, des posts de réseaux sociaux, des Powerpoint, des vidéos, des images...
- Les *données semi-structurées* : elles sont partiellement organisées. Elles sont plus facilement gérables que les données non structurées car elles possèdent des propriétés organisationnelles cohérentes et définies telles que des métadonnées ou des balises sémantiques. Cependant, leur structure n'est pas rigide et elles peuvent toutefois contenir des incohérences ou des variabilités.
- Les *données structurées* : ce sont les données organisées dans un référentiel formaté et qui sont structurées en cellules ou en colonnes. Elles peuvent être générées par des machines mais aussi par des humains. Elles dépendent d'une base de données relationnelle ou d'un schéma et sont donc rigides.

La qualité des données importées va donc dépendre de la fiabilité des sources et de leur structure.

Les types de chargement des données

Concernant le Big data, il y a deux types de chargement des données assez répandus pour traiter rapidement les grands volumes de données. Le choix de l'un d'eux dépend de votre cas d'utilisation.

Le mode Batch (par lots)

Le mode de traitement des données par lots/batch est celui qui est mobilisé traditionnellement pour l'approche ETL : il s'agit de traiter un grand volume de données en une seule fois, sur une période donnée (un traitement des données à la demande). Il faut attendre la fin de la phase de collecte de données (extraction et chargement) pour débuter celle de traitement qui se fait par lots de données (la "*fenêtre de batch*").

Fiche Introduction au Data processing

Il y a donc une période de latence entre le moment où vos données apparaissent dans la couche de stockage et le moment où elles sont disponibles dans les outils d'analyse et de reporting. Vous ne devez que peu intervenir dans cette phase de traitement car les tâches s'exécutent les unes après les autres selon les priorités déterminées et sans interruption. Vous devez juste indiquer le nombre de données à atteindre (la limite) ou le moment précis où l'ETL par lots doit être exécuté (toutes les 24 heures, tous les 3 jours,...).

Le mode *Batch* est souvent utilisé lorsqu'il faut réaliser des calculs complexes qui prennent du temps comme des facturations, des commandes ou encore pour mettre à jour le profil utilisateur d'un site de ventes et proposer des articles personnalisés par exemple.

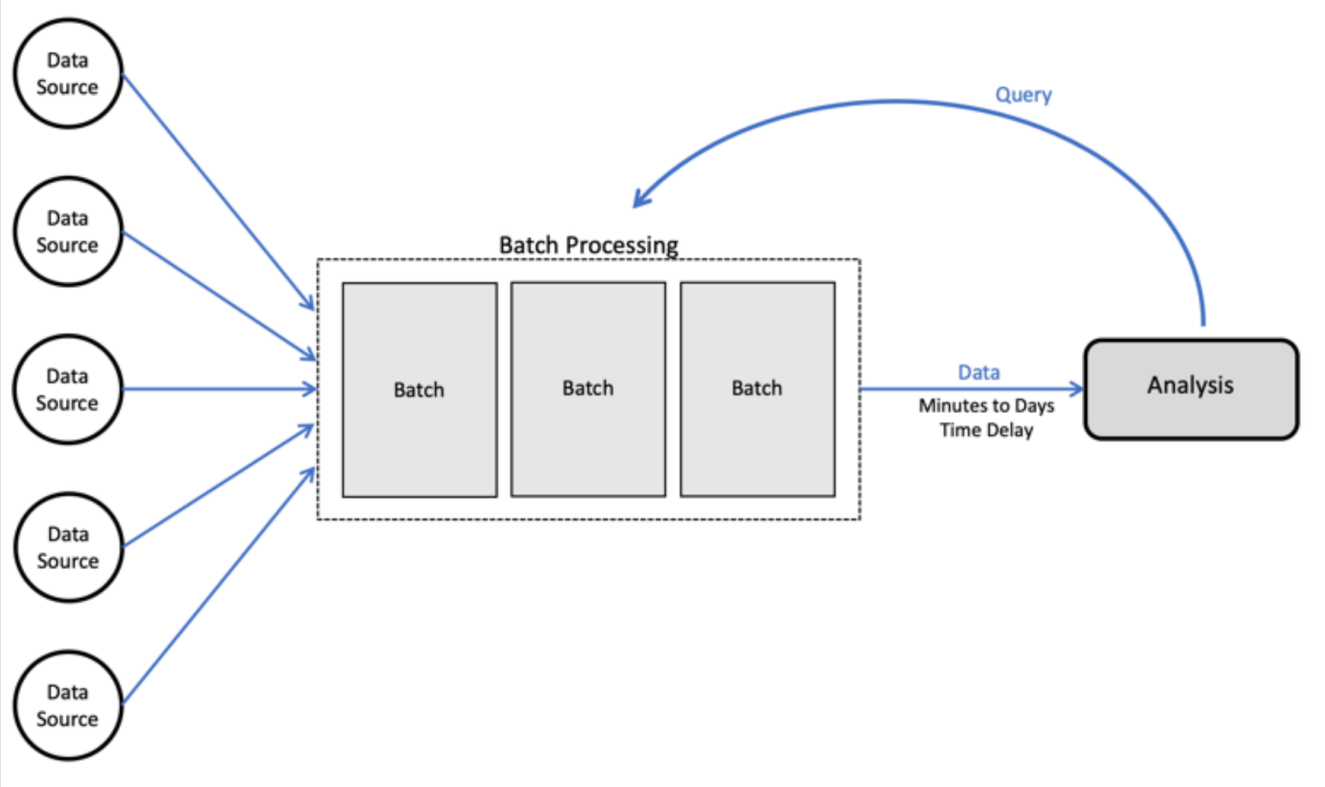


Illustration du mode Batch provenant de Upsolver

Fiche Introduction au Data processing

Le mode Stream

Avec le mode *Stream*, les données sont traitées en continu, c'est-à-dire au fur et à mesure de leur arrivée dans la couche de stockage afin de vous permettre d'accéder rapidement aux données et d'y réagir le plus vite possible une fois un événement détecté. Contrairement au mode batch, le flux est quasi en temps réel et les systèmes ne doivent pas stocker de grands volumes de données.

Le mode *Stream* est surtout conseillé lorsque vous devez détecter des évènements et y répondre rapidement, comme par exemple la surveillance des services, la cybersécurité, l'analyse des comportements, la détection de fraudes, la disponibilité d'un produit en stock, ...

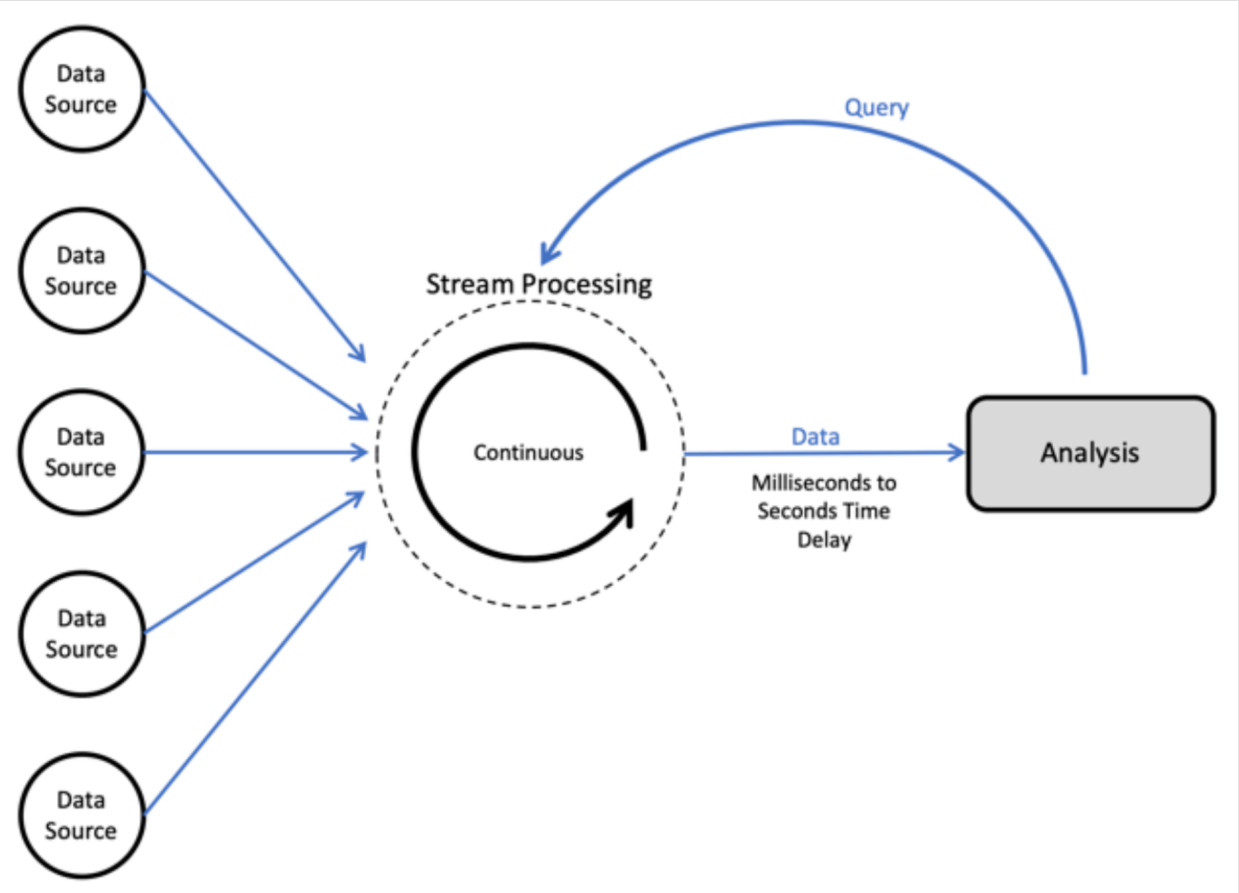


Illustration du mode Stream provenant de Upsolver

Fiche Introduction au Data processing

2 Transformation (*transform*)

La seconde phase concerne le traitement apporté aux données collectées. Il se fait grâce à l'exécution d'algorithmes de machine learning et dépend des utilisations prévues des données (celles-ci doivent être déterminées à l'avance).

Le traitement vise à structurer les données, à les convertir et à les homogénéiser en modifiant leur format, en les enrichissant, en les complétant, en les nettoyant,... Car pour pouvoir croiser, comparer et analyser des données, il faut que celles-ci soient comparables ! Le traitement va donc rendre les données brutes interprétables selon les besoins (les finalités déterminées des données) et les transformer en information commerciale exploitable.

Un stockage intermédiaire et temporaire

Très souvent, les données extraites lors de la première phase sont stockées temporairement dans une zone de transit (staging area). Cette zone se trouve dans l'ETL utilisé et sert d'intermédiaire entre les sources de données et la cible des données (la destination des données). Vous pouvez y faire toutes les manipulations nécessaires à la transformation de vos données.

Les transformations des données

Pour analyser vos données, il faut d'abord les préparer, c'est-à-dire qu'il faut effectuer différentes transformations sur celles-ci en fonction de vos besoins. Plus vos sources de données sont de faible qualité, plus vous aurez des transformations et du nettoyage à effectuer pour éviter d'avoir des erreurs telles que des données manquantes, redondantes, une erreur lexicale et/ou de format,...

Voici quelques exemples de nettoyages et de transformations possibles de vos données pour que celles-ci soient cohérentes, exploitables et fiables :

- Le *mappage des données* : il s'agit de faire correspondre deux modèles de données, les champs des données extraites à ceux associés dans la destination.
- La *vérification* de la cohérence du format des données et leur *conversion* si nécessaire (des unités de mesure, des dates et heures,...)
- La *déduplication* : c'est l'identification et la suppression des enregistrements présents en plusieurs exemplaires (autrement dit les doublons).
- Le *filtrage* : c'est la sélection de certains enregistrements selon des règles.

Fiche Introduction au Data processing

- *Tri* des données par ordre croissant ou décroissant.
- La *jointure* des données : il s'agit de lier des données provenant de sources différentes.
- Le *fractionnement* d'une colonne unique en plusieurs colonnes.
- L'*agrégation* : regroupement de différentes données.
- La *récapitulation* : calculs pour obtenir des valeurs totales.

3 Chargement (*load*)

La phase de chargement est la phase où vos données traitées et structurées sont chargées et stockées dans la destination cible ; un système centralisé qui peut être une base de données, un fichier, un serveur mais qui est bien souvent un datawarehouse (voir section ci-dessous).

Les avantages

En résumé, un ETL vous permet d'avoir :

- une *migration* et une *intégration* automatisées et rapides de grandes quantités de données provenant de systèmes disparates.
- Un *référentiel de données* : les données sont centralisées dans un endroit unique, ce qui vous apporte une meilleure accessibilité à celles-ci.
- Les *transformations complexes* apportées unifient vos données (même format,...) et vous apportent une vue globale de vos ressources. De plus, ces processus de traitement sont réutilisables !
- un *contrôle* sur l'ensemble de vos ressources : vos données sont de meilleure qualité (les traitements les rendent fiables).
- Une *synchronisation* de vos applications et donc une actualisation instantanée de vos données (vous pouvez y accéder en temps réel).
- Vous pouvez tirer *profit* de vos données traitées.

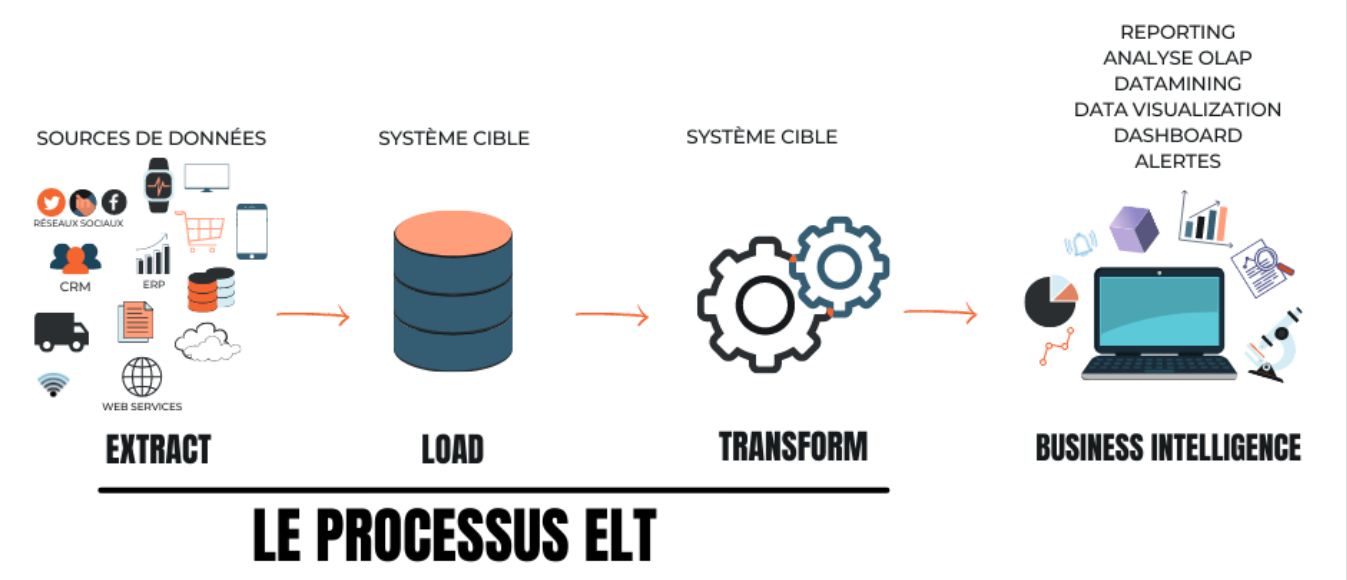
Quelques outils clouds open source

- Talend Open studio
- Scriptella
- Ketl

Fiche Introduction au Data processing

Le processus ELT

Le processus d'intégration *ETL* est traditionnellement utilisé mais avec les nombreuses avancées technologiques, cela évolue. Il existe aussi le processus *ELT* (*Extraction - Loading - Transform*) : dans ce processus, vos données brutes sont extraites des sources de données et sont directement chargées vers votre système cible (très souvent un Data lake, voir la définition dans la section suivante), sans les transformer en fonction de vos besoins métier et sans passer par une zone de transit. Le nettoyage et les transformations des données se font donc après, dans le système/plateforme cible, quand vous devez les utiliser.



Les avantages

En résumé, un *ELT* vous permet d'avoir :

- une *extraction* et un *chargement* plus rapides de vos données dans le système cible vu qu'il n'y a pas toute l'étape de transformation.
- une plus grande *flexibilité* et facilité de *stockage* de nouvelles données non structurées notamment.
- un temps de réflexion plus long concernant la détermination des données à transformer et à analyser.

Fiche Introduction au Data processing

Les modes de stockages et de centralisation de vos données

Il existe deux modes de stockage et de centralisation de données qui sont souvent confondus : les Data lakes et les Data warehouses. Ayant chacun des caractéristiques qui leurs sont propres, le choix d'utiliser l'un ou l'autre va surtout dépendre des besoins spécifiques de votre entreprise.

Les *Data lakes*

Un *Data lake (lac de données)* est un référentiel de données qui vous permet de stocker "en vrac" vos données originales et brutes collectées, et d'y accéder rapidement avant leur traitement (la phase de transformation). Vous pouvez donc y trouver des données *non structurées*, *semi-structurées* ou *structurées*.

En stockant n'importe quel type de données quelles que soient leur nature et leur origine, vous pouvez découvrir de nouvelles questions/problématiques/hypothèses auxquelles vous ne pensiez pas auparavant… C'est d'ailleurs pour cela que les *Data lakes* sont surtout utilisés par des *Data scientists* et *Data engineers* dans le cadre de *Machine learning*, *intelligence artificielle* et *modélisation prédictive*.

L'*ELT* est la solution à privilégier si vous disposez de *lacs de données* car ce processus ingère des données non structurées, contrairement à l'ETL qui transforme les données brutes en données structurées.

Les *Data warehouses*

Les *Data warehouses* sont des bases de données où sont stockées les vues agrégées des données ingérées. Ils ne sont pas connectés aux sources de données directement : ces dernières passent par une solution *ETL* (*Extract*, *Transform and Load*) qui extrait les données pertinentes de sources de données (éventuellement un *data lake*), les transforme et les charge dans le *Data warehouse*.

Les *Data Warehouses* sont surtout utilisés par des analystes, managers et des utilisateurs finaux pour analyser leurs données à l'aide de métriques, reportings et chiffres clés, en vue de prendre plus facilement des décisions.

Fiche Introduction au Data processing

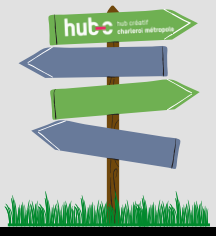
Comparaison entre un Data lake et un Data warehouse

	Data lake	Data warehouse
Données	Brutes Capture de tous types de données Données structurées, semi-structurées, non-structurées Transformations en aval	Historiques de données structurées Traitées (nettoyées et transformées) pour répondre à une finalité précise
Stockage	Très grands volumes Bon marché Lent	Cher Rapide
Flexibilité	Organisation libre Très agile	Schéma fixe Peu agile
Modélisation	Schéma de données à définir en aval mais on peut mettre du tout venant	Schéma de données défini en amont (avant la collecte de données).

Besoin d'une aide supplémentaire ?

Le Hub-C dans le cadre de ses services d'accompagnement numérique organise des workshops et groupes de travail en lien avec les nouvelles technologies de prototypages. Vous souhaitez un accompagnement pour votre projet innovant ou vous souhaitez participer à un prochain workshop ? N'hésitez pas à contacter un membre du Hub !

Vous avez une question spécifique à propos d'une fiche? Elles sont réalisées par les experts du CETIC (Centre d'Excellence en Technologies de l'Information et de la Communication), un centre de recherche appliquée en informatique situé à Charleroi. Vous trouverez toutes les coordonnées ici.



hub créatif charleroi métropole

http://www.hub-charleroi.be

info@hub-charleroi.be

Partenaires

ULB Université Libre de Bruxelles · IGRETEC · CIE Charleroi Entreprendre · Ocetic · MWS enmieux.be · FEDER Union européenne · Wallonie

Le fonds européen de développement régional et la Wallonie investissent dans votre avenir