DATA WAREHOUSE / DATA LAKE

DATA WAREHOUSE

Définition

- un Data Warehouse (entrepôt de données) est une base de données créée pour les analyses de données, la prise de décision et les activités de type Business Intelligence
- un Data Warehouse reçoit des données provenant de sources diverses
- les informations stockées dans un entrepôt sont historisées chronologiquement
- elles offrent une vue d'ensemble de l'évolution d'une information
- les données redondantes existent donc dans un entrepôt de données et offrent aux utilisateurs plusieurs vues de la même information
- exemples : l'évolution des chiffres d'affaires clients sur plusieurs mois ou plusieurs années, l'évolution de l'effectif d'une entreprise ...

DATA WAREHOUSE

Définition

- un Data Warehouse est défini par son sujet
- exemple : un entrepôt de données est créé spécialement pour analyser les données de ventes de l'entreprise et fournit des analyses telles que le palmarès clients, les produits les plus vendus, la répartition géographique des ventes …
- il est également défini par :
  - le type de données qu'il contient
  - les personnes qui utilisent l'entrepôt et analysent ses données
- exemple :
  - type de données : données financières relatives à des clients et des fournisseurs
  - personnes qui utilisent l'entrepôt : le service comptable de l'entreprise

DATA WAREHOUSE

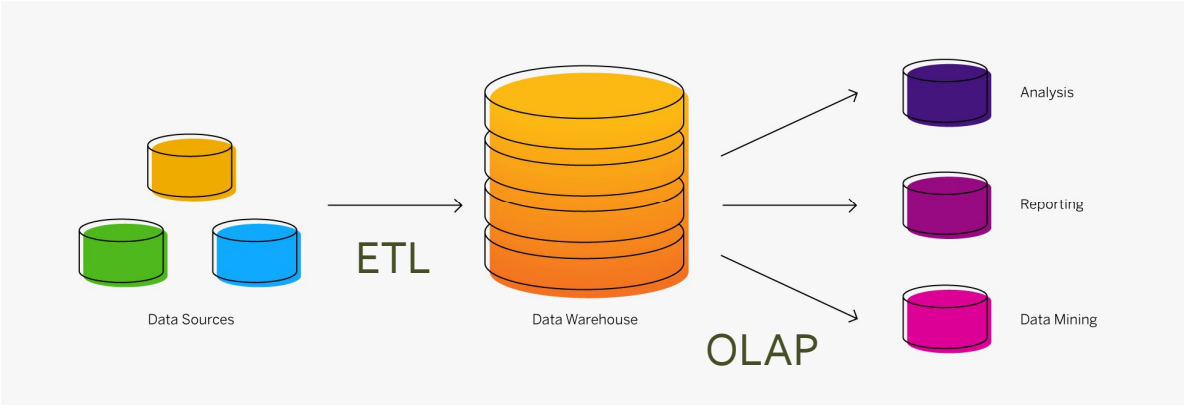
Définition

- un Data Warehouse est non-volatile : une donnée stockée ne doit plus changer
- un Data Warehouse est time-variant : il permet les analyses sur les changements survenus au fil du temps afin de découvrir des tendances
- avec ce stockage de plusieurs versions de la même information, on cherchera à prévoir les prochaines valeurs de cette même information
- exemple : on analyse sur plusieurs années les données de ventes d'un secteur géographique et on essaye de comprendre pourquoi elles ont évolué pour prévoir ce que seront ces mêmes données à l'avenir

DATA WAREHOUSE

Outils

ETL
OLAP



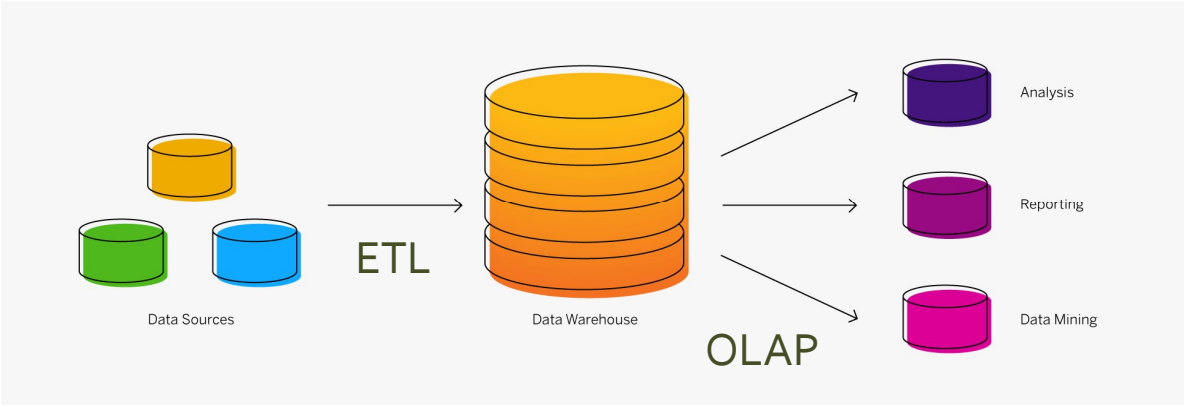
Un environnement Data Warehouse intègre :

- un outil d'extraction, de transformation et de chargement des données appelé ETL (Extract – Transform – Load) qui agrège les informations provenant de différentes sources et les modélise
- un moteur de traitement analytique en ligne appelé OLAP (Online Analytical Processing) qui génère des rapports et favorise l'analyse de type Data Mining (extraction de connaissances)

DATA WAREHOUSE

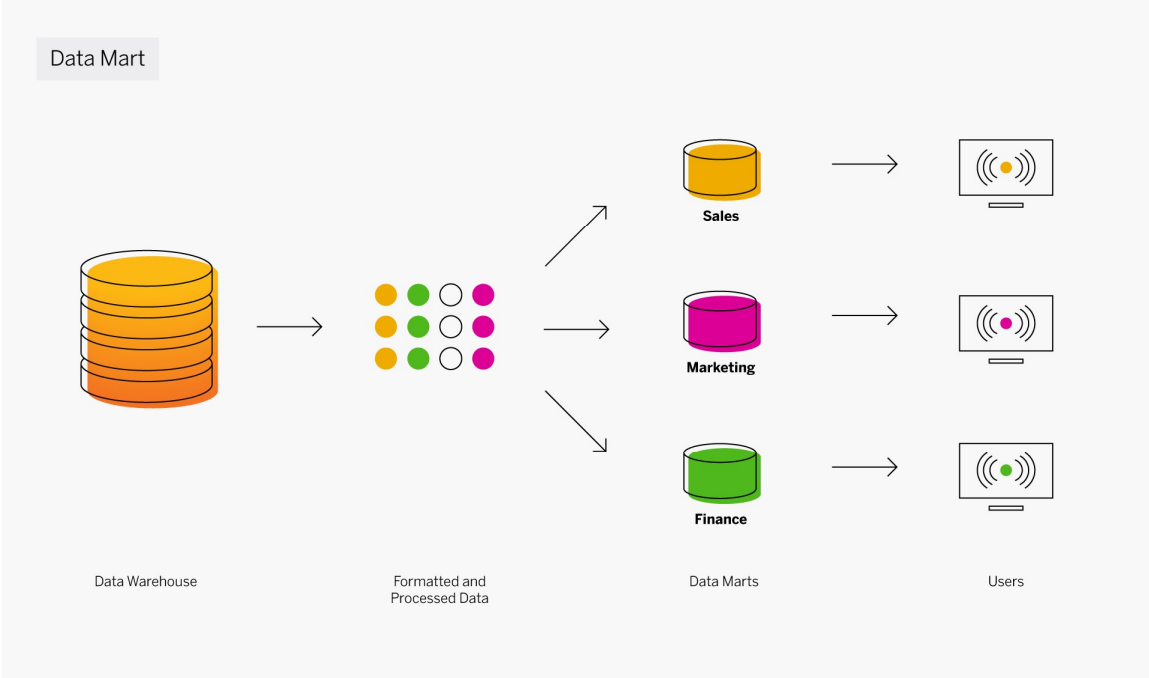
Outils

ETL
OLAP



- grâce à ces outils on résout notamment les conflits de noms et les incohérences en termes d'unités de mesure : c'est ce qu'on appelle l'intégration des données
- exemple :
  - on a on a amassé beaucoup de données contenant des montants dans de nombreuses devises différentes
  - on va les rendre homogènes sur une devise commune

DATA MART



- un Data Warehouse peut être divisé en magasins → Data Marts
- un Data Mart contient des informations dédiées à un domaine spécifique : ventes, marketing, finance …
- ces Data Marts sont analysés séparément

Data Warehouse vs Data Lake

Les 7 différences

1. les données

- un Data Warehouse rassemble une grande quantité de données en provenance de différentes sources de données
- la prise de décision se fait à partir de données déjà rendues cohérentes = données structurées
- un Data Lake est une banque de stockage servant à contenir une immense quantité de données brutes dans leur format d'origine jusqu'à ce que l'entreprise en ait besoin

2. la structure des données

- les Data Warehouses ne peuvent accueillir que des données structurées
- un Data Lake est capable de stocker tous types de données, structurées ou non

Data Warehouse vs Data Lake

Les 7 différences

3. le traitement des données

- les données d'un Data Warehouse sont traitées et passent par l'étape du « schema-on-write » qui leur confère une structure (un modèle) pour devenir exploitables
- les données d'un Data Lake sont stockées sous leur forme brute et ne sont pas exploitables en l'état, et elles ne seront traitées (et donc préalablement mises en forme) qu'en cas de besoin → c'est ce qu'on appelle le « schema-on-read »

4. le stockage

- la différence se fait au niveau des coûts : un Data Lake coûte moins cher qu'un Data Warehouse tant que les données ne sont pas structurées car dans un Data Lake nous n'avons pas besoin de performance de traitement, mais seulement d'un stockage

Data Warehouse vs Data Lake

Les 7 différences

5. l'agilité (évolutions)

- comme un Data Warehouse est structuré, changer sa structure prend du temps
- à l'inverse un Data Lake n'a pas de structure donc les Data Scientists peuvent aisément configurer et reconfigurer les modèles de données, les requêtes et les applications
- le Data Warehouse est donc moins agile, moins flexible

6. la sécurité

- le Data Warehouse est historiquement plus ancien que le Data Lake, ce qui lui confère maturité et fiabilité maximale de sa protection
- la technologie de stockage d'un Data Lake est plus récente donc moins éprouvée techniquement, les données sont donc un peu moins sécurisées au sein d'un Data Lake

Data Warehouse vs Data Lake

Les 7 différences

7. les utilisateurs

- un Data Warehouse est principalement utilisé par les responsables d'entreprises car les données sont prêtes à l'emploi
- un Data Lake est plus utilisé par les Data Scientists car les données sont brutes et en cas de besoin, nécessitent l'expertise de traitement et d'interprétation d'un spécialiste comme un Data Scientist

Data Warehouse vs Data Lake

Les 7 différences

Critères	Data Warehouse	Data Lake
Données	Données cohérentes	Données brutes
Structure des données	Données structurées	Tous types de données
Traitement des données	Directement exploitables	Pas exploitables en l'état
Stockage	Stockage + traitement Coûts plus élevés	Simple stockage Coûts moins élevés
Agilité (évolutions)	Changer la structure prend du temps	Pas de structure donc facilement reconfigurable
Sécurité	Maturité, fiabilité maximale	Technologies plus récentes Moindre fiabilité
Utilisateurs	Décideurs	Data Scientists