Introduction à l'intégration des données et au streaming

Lionel SOUOP

Définition

Intégration des données

L’intégration des données est un ensemble de techniques qui régit le processus d’extraction et de transformation des données, en vue de les rendre exploitables.

Au vu de cette définition assez simpliste, on peut se poser au moins 4 questions:

- Comment extraire les données ?
- Comment les transformer ?
- Quelles sont les implications de “rendre exploitable” ?
- Quelles technologies ?

Définition

Extraction de données

En fonction de l'environnement business dans lequel on évolue, on peut être amené à utiliser des données provenant de différentes sources (un site de e-commerce qui reçoit des données de plusieurs vendeurs).

L'extraction de données consiste à mettre en place des mécanismes de transfert de données d'une plateforme client vers une plateforme propriétaire.

Les moyens de collecte les plus utilisés :

- Via une API
- Via une queue
- Protocole FTP

Définition

Extraction de données

L’extraction de données présente certains défis dont la complexité varie selon l’utilité des données:

- La volumétrie - quel moyen de transfert, quel stockage, quelle rétention, quelles technologies
- La qualité des données - quel niveau de transformation
- Sécurité et confidentialité - Les données sensibles doivent être protégées - anonymisation
- Temps réel vs Batch - choisir la méthode de collecte appropriée

Définition

Transformation des données

Définition et objectifs

- **Conversion et optimisation** : La transformation des données implique de modifier la structure, le format et le contenu des données pour les rendre compatibles avec le système cible ou pour faciliter l'analyse.
- **Amélioration de la qualité des données** : Elle permet de corriger les erreurs, de supprimer les doublons, de normaliser les formats et de garantir la cohérence des données.
- **Préparation pour l'analyse** : Les données transformées sont plus faciles à analyser, ce qui permet d'obtenir des informations plus précises et pertinentes.

Définition

Transformation des données

Principales opérations de transformation

Nettoyage des données :

• Suppression des valeurs manquantes ou incorrectes.

• Correction des erreurs de saisie.

• Suppression des doublons.

Normalisation et standardisation :

• Conversion des données dans un format uniforme (par exemple, dates, devises).

• Application de règles de normalisation pour assurer la cohérence.

Agrégation et regroupement :

• Calcul de statistiques (moyennes, sommes, etc.).

• Regroupement des données par catégories.

Définition

Transformation des données

Transformation de structure :

- Modification de la structure des données (par exemple, conversion de colonnes en lignes).
- Création de nouvelles colonnes à partir de données existantes.

Enrichissement des données :

- Ajout de données provenant de sources externes.
- Création de nouvelles informations à partir des données existantes.

Définition

Transformation des données

Techniques et outils

- ETL (Extract, Transform, Load) : Le processus ETL inclut la transformation des données comme étape intermédiaire entre l'extraction et le chargement.
- ELT (Extract, Load, Transform) : Dans le processus ELT, la transformation a lieu après le chargement des données dans le système cible.
- Outils de transformation de données : De nombreux outils logiciels sont disponibles pour automatiser et faciliter le processus de transformation des données

Définition

Transformation des données

Importance de la transformation des données

- **Amélioration de la prise de décision** : Des données propres et cohérentes permettent d'obtenir des analyses plus fiables.
- **Optimisation des processus métier** : Des données bien structurées facilitent l'automatisation et l'efficacité des opérations.
- **Support des analyses avancées** : La transformation des données est essentielle pour préparer les données pour le machine learning et l'intelligence artificielle.

Définition

Exploitation des données

A la suite de la transformation des données, celles ci doivent être exploitable par des utilisateurs. Pour cela, elle doivent être stockées:

- Dans un Data Lake
- Dans une Data warehouse
- Dans un Delta Lake
- Dans un Datamesh
- Etc …

Bien sur le choix de la technique de stockage dépend des cas d’usages.

Définition

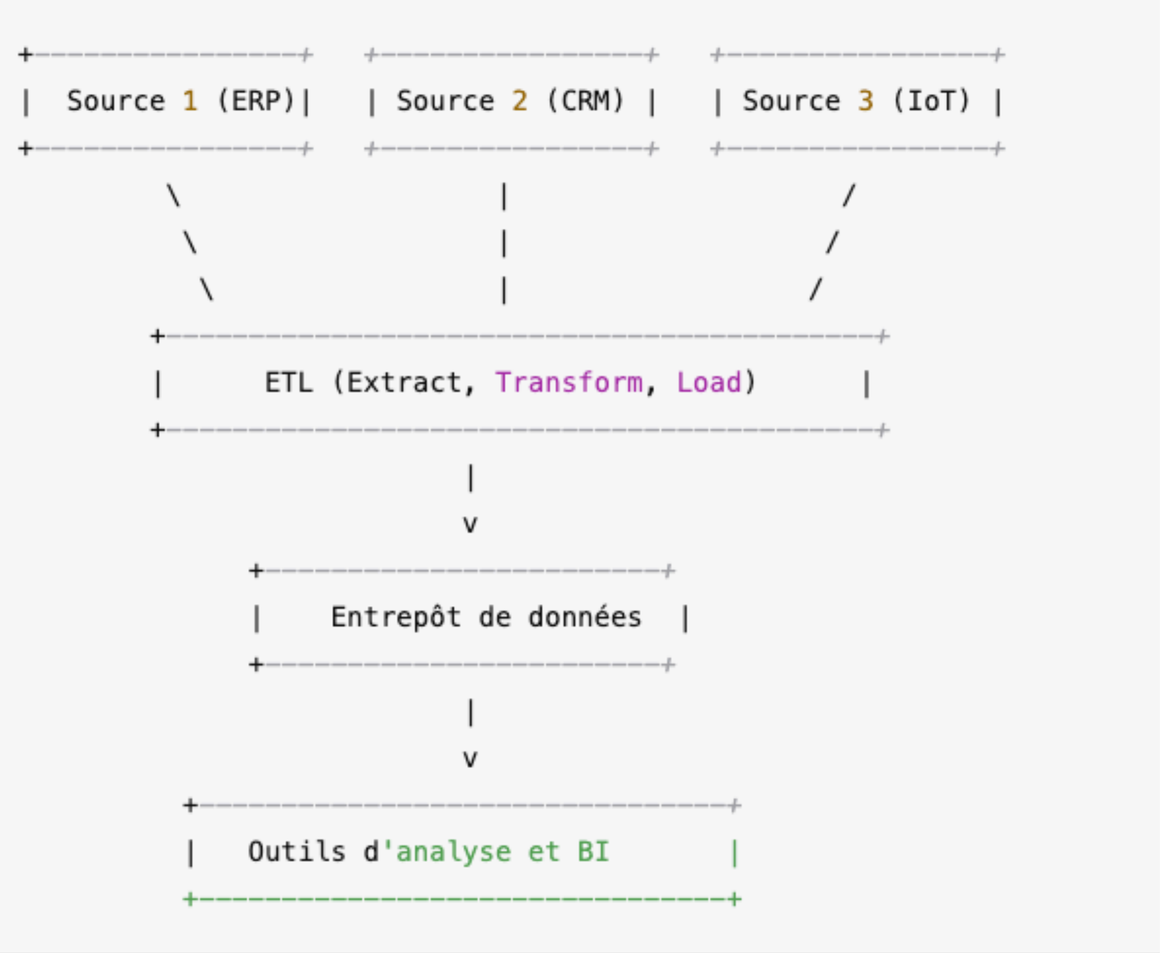
Exploitation des données

Au dessus des moyens de stockage, il est possible de:

- Faire de l’analyse des données
- Exposer des données via une API
- Exposer des données via une doc et des chemins d’accès

Définition

Recap



Intégration des données en streaming

Intégration des données en streaming

Définition

L’intégration des données en streaming est le processus de collecte, de transformation et d’analyse de données en temps réel à partir de flux continus. Contrairement aux méthodes traditionnelles ETL (Extract, Transform, Load) qui traitent les données en batch, le streaming permet de traiter les données au fur et à mesure de leur arrivée.

Intégration des données en streaming

Importance du streaming

- Réduction de la latence : Permet d'obtenir des insights en temps réel.
- Réactivité : Indispensable pour les applications nécessitant des réponses immédiates (ex : surveillance, IoT, finance).
- Traitement en continu : Évite l’attente des traitements batch.
- Scalabilité : Peut traiter des volumes massifs de données provenant de multiples sources.

Intégration des données en streaming

Domaines d'application

- Finance : Détection de fraudes en temps réel.
- IoT (Internet des objets) : Monitoring des capteurs en temps réel.
- Marketing digital : Analyse des interactions utilisateurs en temps réel.
- Cybersécurité : Détection des menaces en temps réel.

Intégration des données en streaming

Batch vs Streaming

Critère	**ETL Batch**	**Streaming**
Fréquence de traitement	Périodique (ex : toutes les heures)	En continu
Latence	Élevée	Faible
Exemples d'outils	Talend, Apache Nifi	Kafka, Flink, Spark Streaming
Cas d'utilisation	Reporting, Data Warehousing	Monitoring, IoT, transactions financières

Les technologies clés

Open source et sur le cloud

Les technologies clés

Open source

- **Les queues**
  - Apache Kafka
  - RabbitMQ
- **Les outils de processing**
  - Apache Spark Streaming - micro batch
  - Apache storm - vrai streaming
  - Apache Flink
- **Stockage**
  - HDFS
  - NoSQL
  - Data Lake/Delta Lake

Les technologies clés

Cloud

- **AWS**
  - Amazon Kinesis - service de streaming permettant l'ingestion, le traitement et l'analyse des flux de données en temps réel.
  - AWS Lambda - Traitement serverless en temps réel des événements.
  - SQS
- **Azure**
  - Azure Event Hubs : Service de streaming haute performance pour collecter et traiter des millions d'événements par seconde.
  - Azure Stream Analytics : Service de traitement de données en temps réel basé sur SQL.
  - Azure Data Explorer (ADX) : Outil puissant d'analyse des flux de données.
- **GCP**
  - Pub/Sub (Publish-Subscribe) : Messagerie asynchrone pour le streaming d'événements.
  - Dataflow : Service de traitement des flux basé sur Apache Beam.
  - BigQuery Streaming : Ingestion en temps réel des données pour analyse immédiate.
- **Confluent Platform**
  - Kafka and KSQLDB