Ressource R5.C.06
Exploitation de la base de données (DATA5)

IUT La Rochelle
BUT INFO3 Développeur IA

M. Sakkari

Objectif

L'objectif de cette ressource est d'exploiter, analyser et évaluer la richesse de données, structurées ou non, appartenant à l'entreprise ou non, pour établir des scénarios permettant de comprendre et d'anticiper de futurs leviers métiers ou opérationnels pour l'entreprise. Cette ressource permettra d'évaluer la qualité et la richesse des données, les analyser et en restituer les résultats pour ensuite les intégrer dans le système d'information cible du métier

Plan

1. **Le Processus Extract-Transform-Load (ETL)**
2. **Exploitation d’un entrepôt de données OLAP**
3. **L’Informatique Décisionnelle**

Evaluation

1. **TPs notés**
2. **Mini-projet**
3. **Examen pratique de fin de formation**

Le Processus Extract-Transform-Load (ETL)

I. Richesse des Données de l'Entreprise

Les entreprises accumulent une variété de données provenant de différentes sources et départements.

1. **Données Structurées**
   - Bases de Données Relationnelles : MySQL, PostgreSQL, Oracle, SQL Server.
   - Feuilles de Calcul : Excel, Google Sheets.
   - ERP et CRM : SAP, Oracle ERP, Salesforce, HubSpot.

I. Richesse des Données de l’Entreprise

2. **Données Semi-Structurées**
   - Fichiers XML et JSON : Données exportées de systèmes internes ou APIs.
   - Logs et Journaux : Fichiers de logs d'applications, journaux de serveur.
   - Emails : Contenus d'emails structurés en partie par des balises ou des métadonnées.

I. Richesse des Données de l'Entreprise

3. **Données Non Structurées**
   - Documents Textuels : Word, PDF, textes libres.
   - Médias : Images, vidéos, fichiers audio.
   - Messages et Discussions : Chats, forums internes.

II. Difficultés

La gestion des données d'entreprise peuvent rencontrer plusieurs défis :

1. **Hétérogénéité des Sources de Données**
   - Formats Variés : Données provenant de bases de données relationnelles, fichiers Excel, documents PDF, etc.
   - Systèmes Divers : Utilisation de multiples systèmes comme ERP, CRM, systèmes de gestion de projet, chacun avec ses propres formats et APIs.

II. Difficultés

2. **Qualité des Données**
   - Données Incomplètes : Absence de certaines informations essentielles dans les enregistrements.
   - Données Erronées : Erreurs humaines, données obsolètes ou incorrectes.
   - Données Dupliquées : Enregistrements en double provenant de différentes sources.

II. Difficultés

4. Intégration et Interopérabilité

- Compatibilité des Systèmes : Difficulté à intégrer des systèmes disparates et à assurer leur interopérabilité.
- Migration des Données : Défis liés à la migration de données entre anciens et nouveaux systèmes sans perte ni corruption de données.

II. Difficultés

5. Big Data

- Volume des données : Gestion de grandes quantités de données nécessitant des infrastructures de stockage et de traitement spécialisées.
- Performance : Assurer la rapidité d'accès et de traitement des données volumineuses.

III. Stratégies pour surmonter les Difficultés

1. Utilisation d'Outils ETL
2. Gouvernance des Données : Politiques de Gestion des Données
3. Utilisation du Cloud et du Big Data : Infrastructure Évolutive, Technologies Big Data

IV. Le Processus Extract-Transform-Load

A. Extract : Extraction depuis des sources de données

L’extraction des données de l'entreprise implique de recueillir des informations provenant de diverses sources internes pour les préparer à des fins d'analyse, de reporting, et de prise de décision.

Les principales considérations et étapes impliquées dans l'extraction des données :

IV. Le Processus Extract-Transform-Load

1. **Identification des Sources de Données :**

- Systèmes ERP (Enterprise Resource Planning) : SAP, Oracle ERP,
- Systèmes CRM (Customer Relationship Management) : Salesforce
- Bases de Données sql/nosql
- Logiciels spécifiques à l'entreprise
- Fichiers Excel, documents Word, PDF

IV. Le Processus Extract-Transform-Load

2. Connexion aux Sources

- Utilisation des APIs et des connecteurs fournis par les systèmes ERP et CRM.
- Utilisation de connecteurs JDBC/ODBC pour un accès direct aux bases de données.
- Utilisation de bibliothèques et d'outils pour lire les fichiers
- Mise en place de mécanismes d'authentification et de sécurité pour accéder aux sources de données sensibles

IV. Le Processus Extract-Transform-Load

3. Extraction des Données

- Requêtes SQL , NoSQL
- Scripts d’Extraction
- Services Cloud
- Utilisation d'outils ETL comme Talend, Informatica, Apache Nifi ou Air Flow pour automatiser et gérer le processus d'extraction.

IV. Le Processus Extract-Transform-Load

B. Transformation des données : pour correspondre à un nouveau schéma/structure, pour les corriger, pour effectuer des calculs …

Ces opérations permettent de transformer les données extraites dans un format uniforme. Les conflits entre les modèles, les schémas et les données sont résolus durant cette phase.

➔ Convertir les données extraites en un format adapté à l'analyse et au reporting.

Voici les aspects clés de cette étape :

IV. Le Processus Extract-Transform-Load

1. **Objectifs de la Transformation des Données :**
    - Nettoyage des Données
    - Normalisation
    - Agrégation
    - Enrichissement
    - Filtrage

IV. Le Processus Extract-Transform-Load

2. Techniques de Transformation des Données

Nettoyage des Données :

- Suppression des Duplications : Identification et suppression des enregistrements en double.
- Traitement des Valeurs Manquantes : Imputation, suppression ou interpolation des données manquantes.
- Correction des Erreurs : Détection et correction des erreurs de saisie, incohérences.

IV. Le Processus Extract-Transform-Load

Normalisation et Conversion :

- Conversion de Formats : Transformation des dates, des devises, et d'autres formats de données en formats standardisés.
- Normalisation des Noms : Uniformisation des noms de champs et des valeurs pour maintenir la cohérence.

IV. Le Processus Extract-Transform-Load

Agrégation et Résumé :

- Calcul de Sommes, Moyennes, Médianes : Regroupement des données par catégories et calcul de statistiques résumées.
- Création de Groupes : Regroupement des données par catégories pertinentes (ex. par région, par produit).

IV. Le Processus Extract-Transform-Load

Enrichissement des Données :

- Ajout de Données Contextuelles : Intégration de données externes (ex. données démographiques, données météorologiques).
- Création de Champs Dérivés : Génération de nouvelles variables à partir des données existantes (ex. ratios, indicateurs de performance).

IV. Le Processus Extract-Transform-Load

Filtrage et Sélection :

- Sélection des Enregistrements : Choix des enregistrements pertinents pour l'analyse (ex. filtrer par date, par région).
- Sélection des Champs : Choix des colonnes pertinentes pour l'analyse.

IV. Le Processus Extract-Transform-Load

3. Outils et Technologies pour la Transformation des Données

- Outils ETL
- Scripts Personnalisés : Python (pandas, numpy), R, Java.
- Services Cloud : AWS Glue, Google Cloud Dataflow, Azure Data Factory.
- Requêtes SQL complexes pour transformer les données directement dans les bases de données

IV. Le Processus Extract-Transform-Load

C. Load – Chargement des Données vers une Cible de Données

L'étape de chargement (Load) dans le processus ETL consiste à déplacer les données transformées vers une destination finale pour l'analyse, le reporting, et l'utilisation opérationnelle.

Cette destination peut être une base de données, un data warehouse, un data lake, ou tout autre système de stockage de données. Voici les aspects clés de cette étape :

IV. Le Processus Extract-Transform-Load

1. **Objectifs du Chargement des Données**
   - Insertion des Données : Ajouter les données transformées dans la cible de données.
   - Mise à Jour des Données Existantes : Actualiser les enregistrements existants avec des données nouvelles ou corrigées.

IV. Le Processus Extract-Transform-Load

2. **Types de Cibles de Données**
   - Bases de Données Relationnelles : MySQL, PostgreSQL, Oracle, SQL Server.
   - Data Warehouses : Amazon Redshift, Google BigQuery, Snowflake.
   - Data Lakes : Amazon S3, Azure Data Lake, Google Cloud Storage.
   - Systèmes NoSQL : MongoDB, Cassandra, HBase.
   - Applications Métier : Systèmes ERP, CRM, outils de BI (Business Intelligence).

IV. Le Processus Extract-Transform-Load

3. Méthodes de Chargement des Données

- Chargement Initial : Processus de chargement massif des données lorsque le système est mis en place pour la première fois.
- Chargement Incrémental : Chargement des nouvelles données ou des données mises à jour depuis le dernier chargement.
- Chargement Complet : Rechargement complet des données à intervalles réguliers.

IV. Le Processus Extract-Transform-Load

- Le chargement par patch implique de charger des mises à jour incrémentielles par lots, souvent appelées "patchs". Cette méthode est couramment utilisée pour mettre à jour les données existantes sans avoir à recharger l'intégralité de la base de données.
- Le chargement par streaming permet de charger les données en continu à mesure qu'elles sont générées ou reçues, plutôt que de les traiter par lots.
- Utilisation d'APIs pour charger les données

Fin

Merci Pour Votre Attention