Mini-projet

Objectif général

L'objectif principal de ce projet est de concevoir et de mettre en œuvre un **système décisionnel** complet permettant de **centraliser** et **analyser** les **données** d'une **entreprise fictive** spécialisée dans la **gestion énergétique des bâtiments intelligents**.

Vous devez **Extraire**, **Transformer** et **Charger** (ETL) des données issues de **plusieurs sources hétérogènes** (fichiers CSV et JSON, base de données relationnelle MySQL), qui nécessitent une consolidation afin d'offrir une vision unifiée, claire et globale pour surveiller la **consommation énergétique** et analyser la **rentabilité** et **l'impact environnemental** de ces bâtiments.

Les données doivent être chargées dans **trois Data Marts** thématiques **indépendants**, chacun dédié à un **axe d'analyse** différent: **1- Consommation énergétique**, **2- Rentabilité économique**, **3- Impact environnemental**.

L'ensemble du **processus ETL** doit être **automatisé et planifié** pour s'exécuter **périodiquement** sans intervention humaine.

Contexte et Objectifs spécifiques :

L'entreprise **GreenCity** gère plusieurs **bâtiments intelligents** répartis sur **différentes régions (sites)**.

Elle souhaite **centraliser ses données** pour :.

- surveiller la **consommation énergétique** (électricité, eau et gaz)
- évaluer la **rentabilité économique** par client, bâtiment ou région (site).
- mesurer l'**impact environnemental** (émissions CO_2, taux de recyclage).
- améliorer la **prise de décision** grâce à des tableaux de bord analytiques.

Portée du projet

- Création d'une **base de données opérationnelle**.
- Génération de fichiers **JSON** et **CSV**.
- Conception et création de **trois Data Marts**.
- Mise en place d'un **processus ETL complet**.
- **Automatisation périodique** du processus **ETL**.
- Réalisation d'un **reporting interactif**.

Les axes d'analyse

1- Consommation énergétique

Mesurer, suivre et comparer la **quantité d'énergie consommée** par client, par bâtiment, par région et par période.

Les **KPI** (les indicateurs de performance) à analyser (liste non exhaustive) :

- Consommation totale (kWh) : par client, par bâtiment, par région et par période.
- Évolution de la consommation dans le temps (Graphique en courbe)
- Consommation vs Température (on cherche une **corrélation)**

2- Rentabilité économique

Analyser la performance financière liée aux bâtiments, clients et Régions.

Les **KPI financiers** à analyser (liste non exhaustive) :

- Chiffre d'affaires (CA) : Somme des montants TTC
- Recouvrement des paiements (taux de paiement)
- Profitabilité / Marge : marge = montant_TTC - coût_énergie
- Rentabilité : par bâtiment, par région, par type d'énergie et par client.
- Comparaison entre clients : classement des clients les plus rentables.

3- Impact environnemental

l'impact environnemental d'un bâtiment correspond à **l'empreinte écologique** de ses activités, mesurée à partir des **émissions de CO_2** (quantité de gaz à effet de serre produite par ses activités) et le **taux de recyclage** (indicateur de gestion des déchets. Plus il est élevé, mieux c'est).

Les principaux **KPI environnementaux** (liste non exhaustive):

- Émissions totales de CO_2 : par bâtiment, par région et par période.
- Évolution des émissions dans le temps : pour voir si un bâtiment (région) pollue plus ou moins qu'avant.
- Comparaison entre bâtiments : classement des bâtiments les plus polluants.
- Analyse du taux de recyclage
- Ratio CO_2 / consommation énergétique : indique si la pollution est proportionnelle ou excessive.

Sources de données opérationnelles

Les sources de données opérationnelles sont celles qui fournissent les données nécessaires à l'alimentation du Data Warehouse. Dans le cadre de ce projet, nous considérons **plusieurs sources hétérogènes** : des fichiers CSV et JSON ainsi qu'une BDR MySQL.

Source	Type	Format
Système IoT	Fichier **JSON**	Données capteurs collectées automatiquement **chaque heure** (par bâtiment par compteur) : consommation
Système de facturation	BD **MySQL**	Clients, Factures, Paiements, Bâtiments, Régions, Clients, Compteurs, etc.
Rapports environnementaux	Fichiers **CSV**	Indicateurs écologiques. Mesures mensuelles d'émissions CO_2 et taux de recyclage.

Structure des fichiers sources

Nous considérons deux types de fichiers : **CSV** et **JSON**

Fichier JSON : Consommation Électricité (ex. Elec_consumption_01_2025.json)

```
{
 "id_region": "REG01"
 {
  "id_batiment": "BAT001",
  "type_energie": "electricite",
  "unite": "KWh",
  "date_generation": "2025-01-14",
  "mesures": [
   {
    "compteur_id": "ELEC_001",
    "date_mesure": "2025-01-14T08:00:00",
    "consommation_kWh": 120.5
   },
   {
    "compteur_id": "ELEC_001",
    "date_mesure": "2025-01-14T09:00:00",
    "consommation_kWh": 123.2
   },
   {
    "compteur_id": "ELEC_002",
    "date_mesure": "2025-01-14T08:00:00",
    "consommation_kWh": 98.1
   },
   {
    "compteur_id": "ELEC_002",
    "date_mesure": "2025-01-14T09:00:00",
    "consommation_kWh": 101.7
   }
  ]
 },
 {
   "id_batiment": "BAT101",
   "type_energie": "electricite",
   "unite": "KWh",
   "date_generation": "2025-01-14",
   "mesures": [
   {
     "compteur_id": "ELEC_101",
```

```
    "date_mesure": "2025-01-14T08:00:00",
    "consommation_kWh": 115.2
   },
   {
    "compteur_id": "ELEC_102",
    "date_mesure": "2025-01-14T08:00:00",
    "consommation_kWh": 45.5
   }
  ]
 }
}
```


Fichier JSON : Consommation Eau (ex. Eau_consumption_01_2025.json**)**


```
{
 "id_region": "REG01"
 {
  "id_batiment": "BAT001",
  "type_energie": "eau",
  "unite": "m3",
  "date_generation": "2025-01-14",
  "mesures": [
   {
    "compteur_id": "EAU_001",
    "date_mesure": "2025-01-14T08:00:00",
    "consommation_m3": 1.5
   },
   {
    "compteur_id": "EAU_001",
    "date_mesure": "2025-01-14T09:00:00",
    "consommation_m3": 1.67
   },
   {
    "compteur_id": "EAU_002",
    "date_mesure": "2025-01-14T08:00:00",
    "consommation_m3": 2.15
   },
   {
    "compteur_id": "EAU_002",
    "date_mesure": "2025-01-14T09:00:00",
    "consommation_m3": 2.23
   }
  ]
 }
```

```
}
```

Fichier JSON : Consommation Eau (ex. Eau_consumption_01_2025.json**)**

```
{
 "id_region": "REG01"
 {
  "id_batiment": "BAT001",
  "type_energie": "gaz",
  "unite": "m3",
  "date_generation": "2025-01-14",
  "mesures": [
   {
    "compteur_id": "GAZ_001",
    "date_mesure": "2025-01-14T08:00:00",
    "consommation_m3": 3.5
   },
   {
    "compteur_id": "GAZ_001",
    "date_mesure": "2025-01-14T09:00:00",
    "consommation_m3": 3.67
   },
   {
    "compteur_id": "GAZ_002",
    "date_mesure": "2025-01-14T08:00:00",
    "consommation_m3": 4.35
   },
   {
    "compteur_id": "GAZ_002",
    "date_mesure": "2025-01-14T09:00:00",
    "consommation_m3": 5.03
   }
  ]
 }
}
```

Fichier CSV – Rapports environnementaux (ex. env_reports_01_2025.csv**)**

id_region	id_batiment	date_rapport	emission_CO2_kg	taux_recyclage
REG01	BAT001	2025-01-31	512	0.67
REG01	BAT002	2025-01-31	430	0.71

Les fichiers CSV et JSON doivent être générés automatiquement à l'aide d'outils tels que Mockaroo, ChatGPT, etc.

Base de données relationnelle MySQL :

Concevoir et créer une base de données relationnelle MySQL intégrant toutes les tables que vous estimez pertinentes pour alimenter les trois Data Marts avec les données nécessaires.

Pour générer un grand volume de données réalistes, vous pouvez utiliser des **générateurs de données** comme **Mockaroo** afin de créer des milliers d'enregistrements simulés.

N.B:

Durant la phase de génération des données des différentes sources opérationnelles, veillez à prendre en considération l'introduction volontaire de quelques défauts de qualité des données tels que :

- *Valeurs manquantes (missing values)*
- *Doublons*
- *Espaces inutiles avant ou après les valeurs de type texte*
- *Format de date incorrect*
- *Valeurs numériques incohérentes*

Ces problèmes doivent être détectés et traités lors de la phase de Transformation.

Data Warehouse

Le Data Warehouse est composé de **trois Data Marts.**

1- **Consommation énergétique**
2- **Rentabilité** économique
3- **Impact environnemental**.

Les trois Data Marts sont basés sur une architecture **dimensionnelle** en **étoile**, avec une table de faits et plusieurs tables de dimensions.

- **Table de faits :** il s'agit de la **table principale** du Data Mart, contenant les **mesures** utilisées pour l'analyse du processus métier. Elle est de type **transactionnel**. Vous pouvez y ajouter toutes les **mesures** jugées pertinentes et cohérentes avec chacun des trois sujets d'analyse.
- **Tables de dimensions :** elles définissent le **contexte d'analyse** et fournissent les détails nécessaires par rapport aux mesures de la table de faits. Vous pouvez ajouter le **nombre** et les **types** de tables de dimensions que vous jugez pertinents pour chacun des trois sujets d'analyse.

N.B :

*Veillez à ajouter des colonnes de type Timestamp (**date d'ajout** et **date de mise à jour** d'une **ligne**) au niveau de la table des faits et les tables de dimension.*

1. Processus ETL

Le processus ETL est l'ensemble des étapes d'Extraction, de Transformation et de Chargement nécessaires pour l'alimentation du Data Warehouse à partir des sources de données opérationnelles. L'ETL est un **processus périodique** dont la **fréquence d'exécution** dépend généralement des besoins métiers.

1.1. Extraction :

Les données nécessaires pour alimenter les trois Data Marts doivent être extraites à partir des différentes sources de données opérationnelles.

Les données extraites doivent être stockées dans des fichiers CSV qui constituent le **Staging Area** (zone de stockage temporaire).

Il est nécessaire de mettre en place une **extraction incrémentale**, ciblant exclusivement les enregistrements non encore extraits ou modifiés. Cela peut se faire à l'aide des **colonnes** de type **Timestamp** (date/heure).

1.2. Transformation :

Les transformations à réaliser sont les traitements nécessaires pour créer un Data Warehouse avec des données **consolidées**, **cohérentes**, **utiles** et **conforme** aux objectifs de ce mini-projet.

N.B :

- *Les données à transformer doivent être extraites à partir des fichiers CSV du Staging Arera.*
- *Après la phase de transformation, les données doivent être sauvegardées dans des fichiers **CSV** dediés.*

1.3. Chargement :

Les données transformées doivent être chargées **périodiquement** dans les Data Marts. Le **chargement périodique** des données doit se faire d'une manière **incrémentale**. En d'autres termes, à chaque nouvelle opération de chargement, seules les nouvelles données opérationnelles ou qui ont subi des mises à jour, doivent être pris en considération.

Outils et Technologies : utiliser l'**ETL Pentaho Data Integration** (PDI) de la suite Pentaho.

1.4. Automatisation du processus ETL :

L'exécution du processus ETL doit être automatisée à une fréquence précise, par exemple chaque jour à 02h00.

Pour Automatiser le processus ETL vous pouvez utiliser le serveur **carte** de Pentaho avec l'outil **Task Scheduler** de Windows (ou **cron** de Linux/MacOS), ou l'outil **Apache Airflow** ou tout autre outil équivalent.

2. Reporting

Le **Reporting** vise à tirer parti des données consolidées afin d'optimiser le processus de prise de décision.

L'objectif principal est de fournir aux utilisateurs un **aperçu global** et des **analyses détaillées** via des **visualisations** de **différents formats** (graphiques, tableaux de bord, rapports, …).

Vous devez créer un **tableau de bord** qui regroupe plusieurs visualisations (graphiques et tableaux), offrant un aperçu global des indicateurs de performances (KPI) les plus importants.

Un **tableau de bord interactif** doit être aussi créé pour permettre une exploration intuitive des données.

Outils et Technologies : Utiliser des outils de Business Intelligence tels que **Pentaho Report Designer**, **Tableau**, **Power BI**, ou **Google Data Studio**.

3. Organisation et déroulement

- Le mini-projet doit être réalisé en **équipe** de **2 à 3 membres**.
- Un **rapport PDF** détaillant toutes les étapes de réalisation du mini-projet, doit être remis le jour de la présentation.
- Les présentations des mini-projets auront lieu le **lundi 05 janvier 2026**.