Etude de cas – Data Management

Consigne : Par groupe de 4 (5 si une personne seule), traiter l'un des quatre sujets ci-dessous.

Evaluation : présentation de 10 minutes qui aura lieu à la dernière heure du module.

Seront pris en compte :

- La clarté dans la présentation du sujet et de la problématique
  La méthodologie (explication des difficultés rencontrées, des choix réalisés)
- La présentation du résultat (KPI) et de la question ouverte (stratégie)
- La réponse aux questions posées par les autres groupes (s'il y en a)

Sujet 1 : RH & "People Analytics" (Focus Sécurité & RGPD)

Scénario :

Suite à une fusion, vous récupérez un fichier CSV mal structuré contenant les données de 15 000 employés. Le DRH souhaite un tableau de bord sur la parité salariale, mais ces données contiennent des informations personnelles critiques (PII).

Données brutes : sujet1_employees_raw.csv

1. Ingestion & Nettoyage

- Chargez le fichier CSV
- Identifiez et supprimez ou corrigez les emails invalides
- La colonne salaire_brut contient du texte ("EUR", "50000 €"). Convertissez tout en *integer*.
- Harmonisez date_embauche au format ISO (YYYY-MM-DD). Rejetez les dates incohérentes.

2. Sécurité & Anonymisation

- Le dashboard final ne doit pas afficher nom et prenom. Remplacez ces champs par un Hash_ID unique.
- La colonne secu_sociale est trop sensible. Masquez tout sauf les 2 derniers chiffres (ex: **********89).
- Créez une fonction qui retourne une version du Dataset filtrée selon le rôle :
  - Role="Admin" : Voit tout.
  - Role="Manager" : Ne voit pas la colonne secu_sociale et les salaires sont arrondis au millier près.

3. Métadonnées & Cycle de vie

- Produisez un tableau (ou YAML) décrivant votre dataset final Gold. Pour chaque colonne (Hash_ID, Salaire, Email...), précisez : *Description, Type, Sensibilité (PII/Confidentiel/Public), Owner*.
- **Backfill & Maintenance :**

Mise en situation : Un bug est découvert 6 mois plus tard : les salaires des personnes embauchés après 2022 sont faux (non prise en compte des augmentations entre 2022 et 2025).

Question : Expliquez comment vous relancez le calcul uniquement sur les mois concernés sans tout recalculer depuis 2018. Quel mécanisme de stockage (format de fichier) faciliterait cette opération ?

4. Visualisation Finale

Produisez un visuel (outil de votre choix) répondant à la question du DRH :

- **Graphique :** Histogramme de la distribution des salaires par sexe, csp et âge ?
- **KPI :** Salaire moyen global et nombre d'emails corrigés/rejetés.

5. Stratégie & Gouvernance

La loi impose de conserver les bulletins de paie pendant une longue durée, mais le RGPD impose de minimiser les données personnelles. Si un employé quitte l'entreprise, décrivez votre **politique d'archivage** : À quel moment passez-vous sa donnée de la base active (Hot storage) à une base d'archive (Cold storage) ? Que supprimez-vous définitivement ?

Sujet 2 : Santé & Épidémiologie (Focus Qualité & Éthique)

Scénario :

Un réseau de cliniques vous transmet les admissions de patients pour suivre les épidémies respiratoires. Les données sont saisies manuellement par des médecins pressés.

Données brutes : patient_admissions.csv

1. Ingestion & Contrôle Qualité

- Chargez le fichier CSV.
- Supprimez les lignes où l'âge du patient est impossible ou la date d'admission anormale.
- La colonne *diag_code* contient parfois des commentaires. Extrayez uniquement le code alphanumérique (ex: "J10.1").

2. Anonymisation Avancée

Pour le partage public, l'anonymat doit être garanti.

- Transformez la date de naissance précise en "Tranche d'âge" (ex: 20-30).

- Transformez le code postal ou l'adresse en conservant uniquement le département (2 premiers chiffres). Supprimez le nom.

3. Métadonnées & Cycle de vie

- Produisez le dictionnaire des données du dataset Gold. Précisez bien quelles colonnes ont subi une transformation pour l'anonymisation.
- **Backfill & Maintenance :**

Mise en situation : La classification des maladies change (le code "J10.1" devient "GRIPPE-A"). Vous devez mettre à jour l'historique pour que les graphiques restent cohérents sur 5 ans.

Question : Décrivez la procédure technique pour appliquer ce "mapping" rétroactivement. Est-il préférable de modifier la donnée brute ou d'appliquer la transformation à la volée ?

4. Visualisation Finale

Produisez une carte ou un graphique :

- **Graphique :** Évolution temporelle du nombre d'admissions pour "Grippe".
- **KPI :** Nombre total d'admissions valides après nettoyage.

5. Stratégie & Gouvernance

Vous stockez des données sensibles. Un patient exerce son droit à l'effacement. Cependant, vous avez besoin de son dossier pour des statistiques de santé publique (anonymes). Quelle solution privilégieriez-vous ?

Sujet 3 : Banque & Fraude (Focus Performance & Intégrité)

Scénario :

Vous devez traiter les logs de transactions bancaires pour détecter les fraudes. Le volume de données est voué à exploser.

Données brutes : transactions_log.csv et blacklisted_ips.csv

1. Ingestion & Typage

- Chargez les fichiers.
- Nettoyez et convertissez en *Float* la colonne *amount*.
- Le système source envoie parfois deux fois la même transaction. Identifiez et supprimez les doublons parfaits.

2. Enrichissement & Règles

- Joignez les transactions avec la liste des *blacklisted_ips*.
    - Créez une colonne is_suspicious qui vaut TRUE si L'IP est blacklistée OU le montant > 5000 €.

- Sauvegardez le résultat final au format **Parquet** avec un partitionnement par année.

3. Métadonnées & Cycle de vie

- Produisez la fiche catalogue du Dataset. Indiquez clairement qui est le "Data Steward" responsable de la liste des IPs blacklistées.
- **Backfill & Maintenance :**

Mise en situation : On s'aperçoit qu'une IP a été ajoutée à la blacklist avec 3 jours de retard. Les transactions de ces 3 jours n'ont pas été marquées "suspectes".

Question : Comment réconcilier cet historique ? Faut-il rejouer le pipeline de détection sur les 3 derniers jours ? Si oui, comment s'assurer qu'on ne crée pas de doublons ?

4. Visualisation Finale

Produisez un rapport pour le service Fraude :

- **Graphique :** Pie chart "Transactions Suspectes vs Normales".
- **Tableau :** Top 5 des montants les plus élevés flaggués comme suspects.

5. Stratégie & Gouvernance

Votre pipeline charge les transactions tous les jours.

1. Quelle est la différence entre une **sauvegarde** de la base de données et une **réplication** ?
2. Si un collègue supprime accidentellement la table Gold (DROP TABLE), laquelle de ces deux solutions permet de récupérer les données ?

Sujet 4 : E-commerce & Client 360 (Focus Réconciliation)

Scénario :

L'entreprise veut unifier ses bases clients "Web" et "Magasin" qui sont gérées séparément. Les données sont très hétérogènes et les doublons nombreux.

Données brutes : web_orders.csv , store_loyalty.csv et conversion_rates.csv

1. Ingestion & Standardisation

- Chargez les deux fichiers.
- Normalisez les numéros de téléphone au format international +33XXXXXXXXX.
- Remettez les noms propres au format "Titre" (Première lettre majuscule, le reste en minuscule) pour faciliter la comparaison.
- Dans web_orders, standardisez tous les montants en Euros (€).

2. Réconciliation

- Tentez de fusionner les deux bases.
- Définissez une règle pour dire "C'est le même client"
- Créez une table finale unique avec une seule ligne par client physique ;
    - Calculez les attributs : Total_Achats_EUR (Web) et Total_Points (Magasin).

3. Métadonnées & Cycle de vie

- Documentez la règle de "Matching" utilisée. C'est une métadonnée métier cruciale ("Business Rule").
- **Backfill & Maintenance :**

Les taux de change fluctuent chaque jour. Vous avez utilisé un fichier de taux fixe.

Si on vous demande de recalculer le CA Web exact avec le taux *du jour de l'achat*, comment modifiez-vous votre pipeline ? Quelles données supplémentaires vous manquent-ils ?

4. Visualisation Finale

- **Graphique :** Nuage de points (Scatter Plot) : *Montant Achats Web* (Axe X) vs *Points Fidélité* (Axe Y). *Objectif :* Identifier les clients "Omnicanaux" (ceux qui sont hauts sur les deux axes).
- **KPI :** Pourcentage de clients identifiés uniquement par leur téléphone (c'est-à-dire sans email).

5. Stratégie & Gouvernance

Votre base de données client est critique pour le site web. Décrivez une stratégie de sauvegarde (fréquence, type de stockage) qui permet de limiter la perte de données à **maximum 1 heure** en cas de crash serveur (RPO - Recovery Point Objective)

Votre base client unifiée devient critique pour le site web. Décrivez une stratégie de sauvegarde (RPO = Recovery Point Objective < 1h) pour éviter de tout perdre en cas de crash.