**Client Data Quality BRD**

Projet Contrôle Qualité des Données Clients

*Prepared by: Azedine Zaaraoui*

# Contexte

L’entreprise XYZ collecte quotidiennement des fichiers CSV contenant des informations sur ses clients. Ces données sont exploitées par les équipes marketing, finance et service client. La qualité des données est essentielle pour garantir la fiabilité des analyses, des campagnes marketing et des opérations de service client.

# Objectifs du Projet

• Mettre en place un pipeline automatisé pour la vérification et le suivi de la qualité des données clients.

• Appliquer des règles de contrôle de qualité pour détecter les erreurs, incohérences et doublons.

• Produire des rapports quotidiens de qualité des données.

• Intégrer des tests automatiques et un processus CI/CD pour assurer la régularité et la robustesse des contrôles.

# Exigences Fonctionnelles

• Le pipeline doit s'exécuter automatiquement à la réception d’un nouveau fichier CSV.

• Les règles de contrôle doivent vérifier la complétude, les doublons, les formats et la cohérence des âges.

• Un rapport de qualité doit être généré pour chaque exécution.

• Les erreurs critiques doivent être signalées (plus de 20% de données invalides).

• Des tests unitaires doivent être exécutés automatiquement via un pipeline CI/CD.

# Règles de Contrôle de Données

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Règle de Contrôle** | **Seuil d’Acceptation** | **Action si Erreur** |
| RC1 | Les colonnes id\_client, email, telephone, date\_naissance doivent être non nulles. | 100% complétude | Identifier les lignes invalides |
| RC2 | Aucun doublon ne doit exister sur la colonne email. | 0% de doublons acceptés | Identifier les doublons |
| RC3 | Aucun doublon ne doit exister sur la colonne telephone. | 0% de doublons acceptés | Identifier les doublons |
| RC4 | Les emails doivent être au format valide (exemple@domaine.com). | 100% format valide | Identifier les emails invalides |
| RC5 | Les numéros de téléphone doivent être au format international (+ indicatif). | 100% format valide | Identifier les téléphones invalides |
| RC6 | L’âge des clients doit être compris entre 18 et 100 ans. | 100% âges valides | Identifier les âges incohérents |

# KPI de Mesure de la Qualité des Données

|  |  |  |
| --- | --- | --- |
| **KPI** | **Description** | **Formule / Exemple** |
| Taux de Complétude | Pourcentage de valeurs non nulles dans les colonnes critiques | $(1 - \frac{Valeurs\ nulles}{Total\ valeurs}) \times 100$ |
| Taux de Validité | Pourcentage de valeurs respectant le format attendu (email, téléphone...) | $(1 - \frac{Valeurs\ invalides}{Total\ lignes}) \times 100$ |
| Taux d’Unicité | Pourcentage de valeurs uniques (sans doublons) | $(1 - \frac{Doublons}{Total\ lignes}) \times 100$ |
| Taux de Cohérence | Pourcentage de données cohérentes selon les règles métiers (ex : âge) | $(1 - \frac{Incohérences}{Total\ lignes}) \times 100$ |
| Volume d’Erreurs | Nombre d’erreurs par catégorie (nulles, doublons, formats, cohérence) | - |
| Score Qualité Global | Moyenne des taux ci-dessus pour un indicateur global de qualité | Moyenne simple des taux de complétude, validité, unicité, cohérence |

# Exemple de Jeu de Données Initial

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **id\_client** | **nom** | **prenom** | **email** | **telephone** | **pays** | **date\_naissance** |
| 1 | Durand | Jean | jean.durand@gmail.com | +33612345678 | France | 1990-05-15 |
| 2 | Martin | Claire | claire.martin@ | 0612345678 | France | 2010-07-20 |
| 3 | Doe | John | john.doe@gmail.com | +33612345678 | USA | 1985-12-01 |
| 4 | Dupont | Sophie | sophie.dupont@gmail | 0601020304 | France | 1965-04-30 |
| 5 | Smith | Anna | anna.smith@gmail.com | +447911123456 | UK | 2005-11-02 |

# Evolution Possible

• Intégration de corrections automatiques pour les erreurs de format simples.

• Connexion à des API externes pour validation des emails et numéros de téléphone.

• Ajout d’un tableau de bord Power BI ou Grafana pour le suivi de la qualité des données.

• Mise en place de notifications automatiques (Slack, email) en cas de seuil critique d’erreur.

# CI/CD et Tests Automatisés

• Mise en place de GitHub Actions pour l’exécution automatique des tests unitaires Pytest à chaque push.

• Validation continue du pipeline Airflow.