Cahier de charge

Détecter les anomalies des données clientèle de la banque

Walid Elaoun

Attijari Bank Tunis

Sommaire

1. Définition Du projet
2. Data gouvernance
3. Data Quality
4. Créations de données source brutes fictives (chargement d'un ODS)
5. Les règles de gestion
6. La chaine décisionnelle
7. Les outils utilisés
8. Définition Du projet :

Ce projet vise à appliquer le processus ETL qui consiste à l’extraction de données provenant de différents fichiers Excel, faire les transformations nécessaires de ces données et les chargées dans la base de données à travers l’outils d’intégration Talend, afin de les exploiter dans la mise en place de mesures et KPIs qui aident à détecter des anomalies et améliorer la qualité des données.

1. Data gouvernance :

La gouvernance des données correspond à l'ensemble des organisations et procédures mises en place au sein d'une entreprise pour encadrer la collecte des données et leurs utilisations. Il s'agit à la fois de respecter les obligations légales imposées par les pays et d'établir des cadres internes pour optimiser l'utilisation des données. Il ne s'agit pas d'appliquer la gouvernance des données à l'échelle du produit ou du service, mais à la dimension de l'entreprise.

**-La disponibilité des données :** Elles doivent être sauvegardées dans le système informatique et être mises à disposition des utilisateurs. Cela implique de donner un accès sécurisé à ces derniers et de leur permettre de modifier ou non les informations tout en minimisant les possibilités d’erreurs. Cette gestion des autorisations s’accompagne d’une copie des données qui permettront en cas de perte ou de dysfonctionnement du service de maintenir la qualité d’un service. Il faut également prévenir les utilisateurs de ces copies.

**-L’utilisabilité des données :** la Data Governance détermine “le degré selon lequel un produit peut être utilisé, par des utilisateurs identifiés, pour atteindre des buts définis avec efficacité, efficience et satisfaction, dans un contexte d’utilisation spécifié”. Défini par Jakob Nielsen (un expert dans le domaine de l’ergonomie informatique), ce concept implique l’efficience, la satisfaction, mais aussi la facilité d’apprentissage et la facilité d’appropriation d’un outil informatique impliquant des terminaux à écrans.

**-L’intégrité des données :** Cette notion est à corréler avec la disponibilité. Elle implique de ne pas altérer ou de détruire les données en assurant qu’elles soient cohérentes, fiables, pertinentes et valides. Il s’agit également de vérifier la qualité de l’écriture sur un support de stockage, la transmission, etc.

**-La sécurité des données :** la Data Governance implique de mettre en place une politique de sécurisation des informations de l’entreprise et de ses clients. Elle doit permettre d’assurer la continuité des trois notions précédentes. Selon la CNIL, il convient donc de la mettre en place à toutes les échelles : contrôler les accès, tracer l’activité et gérer les violations de données, réduire les vulnérabilités, anonymisation, etc.

1. Data Quality :

Garantir des informations précises et durables est la clé de la qualité des données d'une entreprise. Une entreprise doit suivre ses données et mettre en place des indicateurs faciles à communiquer, abordables et complets. La mise en place de ces indicateurs est une étape importante pour toute entreprise souhaitant maîtriser la qualité de ses données.

Les données sont considérées comme étant de qualité lorsqu'elles répondent à certaines normes pour l'usage auquel elles sont destinées. Les données avec des caractéristiques de qualité répondent à des normes spécifiques qui les définissent. On dit que les données dont la qualité des données présente l'une des caractéristiques suivantes : exactitude, exhaustivité, cohérence et précision.

1. Créations de données source brutes fictives (chargement d'un ODS) :

-ODS: Un Operational Data Store (ou ODS) est une structure intermédiaire de gestion des données. Il permet de stocker et de centraliser des données de sources hétérogènes afin de les soumettre plus tard à un traitement par des outils.

1. Adresse.xls :

C’est un fichier Excel contenant les adresses, villes, régions et pays des différents clients de la banque.

1. Nom.xls :

C’est un fichier Excel contenant les noms des différents clients de la banque.

1. Prenom.xls :

C’est un fichier Excel contenant les prénoms des différents clients de la banque.

1. Adresse.xls :

C’est un fichier Excel contenant les adresses des différents clients de la banque.

1. Email.xls :

C’est un fichier Excel contenant les Emails des différents clients de la banque.

1. Tel.xls :

C’est un fichier Excel contenant les numéros de téléphone des différents clients de la banque.

1. CIN.xls :

C’est un fichier Excel contenant les numéros de CIN des différents clients de la banque.

1. Age.xls :

C’est un fichier Excel contenant les dates de naissance des différents clients de la banque.

1. Agence.xls :

C’est un fichier Excel contenant les emplacements des différentes agences de la banque.

Remarque :

Tous ces fichiers contiennent des données fictives qui seront créer au début du processus de travail.

1. Les règles de gestion :
2. Adresse : Contient au moins 3 mots, 1 chiffre et un mots de cette liste : [Rue,Immeuble,Imm,Cité,Route,Avenue,AV,N…..]. Chaque mot doit inclure au moins 2 lettres.
3. Age : Valeur entre 0 et 100
4. CIN : Premier chiffre 0 ou 1 avec exactement 8 chiffres
5. Email : Format standard des adresses mail
6. Téléphone : Format standard des numéros de téléphones tunisiens entre fiabilisé et portable.
7. Nom et prénom : Valeur différente de NULL
8. La chaine décisionnelle :
9. Création d’une base fictive :

Pour garder la confidentialité des données, nous devons créer une base fictive dans oracle pour stocker les données provenant des différents fichiers Excel

1. Application des règles de gestion :

Talend est l’outil idéal pour appliquer les règles de gestion. Deux fichiers seront comme output pour chaque fichier Excel mis en input.

Le premier output correspond aux données incorrectes (anomalies détectées), le deuxième est celui des données correctes.

1. Connexion entre oracle et Power BI :

Afin de pouvoir intégrer les données stockées dans les bases de données il faut avoir une connexion entre Oracle et Power BI.

1. Visualisations dans Power BI :

Les visualisations dans Power BI aident à la prise de décision.

Nous allons manipuler les données afin d’avoir un rapport contenant des KPIs, des graphiques et des mesures utiles pour la détection des anomalies par gestionnaire ou encore par agence.

Exemple de quelques mesures :

1) Nombre d’anomalie par :

-Agence

-Groupe

-Zone

-Gestionnaire

2) Le Client ayant le plus de nombre d’anomalie

3) L’agence ayant le plus de nombre d’anomalie

4) Les valeur d’anomalie les plus répétée

5) Taux d’anomalie dans les adresses

6) Le gestionnaire ayant fait le plus d’anomalies par agence

7) Top 3 des valeurs d’anomalie par agence etc….

1. Les outils utilisés :

* Talend Studio : Processus ETL
* Oracle 11 G : Création des tables dans la base de données
* Power BI Desktop: Creation des visualisation