Paris School of Business

M2 Data Management
Groupe 7

SPRINT 3

Thomas Dahrouj - Nadir Boudjerida - Jean-louis Hu

I. Analyse préliminaire des données : identification des problèmes et incohérences

cf la colonne “ Critères de qualité de données”:

BDD compréhension

II. Établir des indicateurs clés pour le suivi continu de la qualité des données

L'évaluation de la qualité des données est essentielle dans tout projet impliquant leur collecte, leur traitement ou leur analyse. Plusieurs métriques permettent de mesurer cette qualité, chacune offrant un éclairage différent sur les aspects fondamentaux tels que la complétude, la validité, la cohérence ou encore le volume. Cette analyse systématique garantit la fiabilité des conclusions tirées à partir de ces données. Nous explorerons six indicateurs principaux de qualité des données : le taux de complétude, le taux de duplication, le taux de validité, le taux de cohérence, la latence des données, et le taux d'unicité des clés primaires.

1) Le taux de complétude des données

La complétude représente la proportion de champs remplis dans un ensemble de données. Cet indicateur est calculé à l'aide de la formule suivante :

$$\text{Complétude (\%)} = \left(1 - \frac{\text{Nombre de valeurs manquantes}}{\text{Total de champs}}\right) \times 100$$

Cette métrique est cruciale pour identifier les champs ou lignes contenant des valeurs manquantes qui pourraient compromettre l'analyse. Un faible taux de complétude peut indiquer des problèmes dans la collecte des données ou nécessiter des actions de remplacement ou d'imputation.

2) Le taux de duplication

Le taux de duplication mesure la proportion de lignes en double dans un ensemble de données. La formule utilisée est :

$$\text{Duplication (\%)} = \frac{\text{Nombre de lignes dupliquées}}{\text{Nombre total de lignes}} \times 100$$

Cet indicateur aide à surveiller les données redondantes, qui peuvent biaiser les analyses et réduire leur pertinence. Une forte duplication peut indiquer des erreurs dans le pipeline de traitement des données ou la présence de doublons non désirés.

3) Le taux de validité des données

La validité des données correspond au pourcentage de valeurs conformes à des règles définies, telles que le respect d'un format spécifique ou des plages de valeurs prédéterminées. Par exemple, une colonne de dates pourrait exiger le format `YYYY-MM-DD` ou une plage d'années entre 2000 et 2024. La formule est la suivante :

$$\text{Validité (\%)} = \frac{\text{Nombre de valeurs valides}}{\text{Nombre total de valeurs}} \times 100$$

Cet indicateur garantit la conformité des données aux exigences métiers ou techniques, réduisant ainsi les risques d'erreurs dans les analyses.

4) Le taux de cohérence des données

La cohérence des données mesure la proportion d'enregistrements respectant des relations ou contraintes définies entre colonnes ou tables. Par exemple, une clé étrangère (comme site_id) doit exister dans la table correspondante, et une somme totale ne peut pas être négative. La formule utilisée est :

$$\text{Cohérence (\%)} = \frac{\text{Nombre d'éléments cohérents}}{\text{Total d'éléments}} \times 100$$

Un faible taux de cohérence peut révéler des erreurs structurelles dans les données, nécessitant des ajustements au niveau des relations ou des règles métiers.

5) La latence des données

La latence mesure le temps écoulé entre la collecte des données et leur disponibilité pour analyse. Cet indicateur est calculé ainsi :

$$\text{Latence (en heures)} = \text{Horodatage actuel} - \text{Dernier horodatage des données collectées}$$

Un délai excessif dans la mise à jour des données peut indiquer des inefficacités dans le pipeline, compromettant la pertinence des analyses, notamment dans les systèmes en temps réel.

6) Le taux d'unicité des clés primaires

L'unicité des clés primaires représente la proportion de valeurs uniques pour une clé ou un identifiant dans un ensemble de données. Sa formule est :

$$\text{Unicité (\%)} = \frac{\text{Nombre de clés uniques}}{\text{Nombre total de clés}} \times 100$$

Un faible taux d'unicité peut signaler des collisions d'identifiants, ce qui pourrait entraîner des incohérences dans les bases de données relationnelles.

7) Volume de données

Le volume de données correspond au nombre total de lignes ou d'enregistrements présents dans une table ou un ensemble de données. Cet indicateur simple mais significatif permet de surveiller l'intégrité des flux de données. En effet, des variations soudaines ou inattendues du volume peuvent signaler des problèmes dans le pipeline, tels qu'un arrêt inattendu, des erreurs d'extraction ou des doublons non détectés.

Les indicateurs de qualité des données tels que le taux de complétude, de duplication, de validité, de cohérence, de latence et d'unicité offrent une vue complète sur l'état des données utilisées dans un projet. Ils permettent de détecter et de corriger des anomalies avant qu'elles n'affectent les résultats finaux. Une gestion rigoureuse de ces métriques est essentielle pour garantir des analyses fiables et exploitables.

III. Contrôler la qualité des données

La qualité des données est un enjeu majeur dans tout projet analytique, car elle conditionne la fiabilité des résultats obtenus. Pour garantir cette qualité, plusieurs critères clés doivent être respectés, tels que l'unicité, la complétude, la validité, l'exactitude et la consistance. Ces dimensions permettent d'évaluer si les données sont adaptées à une utilisation opérationnelle ou stratégique.

1) Les critères fondamentaux de la qualité des données

Les données doivent répondre à des normes précises pour être considérées comme fiables

- **Unicité**: Les identifiants tels que `id` ou `session_id` doivent être uniques pour éviter les ambiguïtés.
- **Complétude**: Les champs critiques, comme les adresses email ou les horodatages, ne doivent pas contenir de valeurs manquantes susceptibles de nuire à l'analyse.
- **Validité**: Les données doivent respecter des formats spécifiques, par exemple les emails ou les dates conformes aux formats attendus.
- **Exactitude**: Les champs interdépendants doivent rester cohérents entre eux.
- **Consistance**: Les valeurs doivent respecter des règles métier, comme des plages de valeurs raisonnables pour des données temporelles.

Ces critères sont ensuite appliqués à l'ensemble des tables pour en évaluer la qualité globale.

L'analyse a été menée sur plusieurs tables, chacune présentant des particularités en termes de qualité des données.

A. Table `owa_visitor`

- **Valeurs manquantes** : Très nombreuses (41 920), nécessitant une attention particulière.
- **Doublons** : Aucun détecté.
- **Valeurs aberrantes** : Quelques incohérences détectées dans les colonnes liées aux données temporelles (ID, année).
- **Recommandation** : Identifier les colonnes spécifiques contenant des valeurs manquantes et corriger les aberrations temporelles pour améliorer la cohérence.

B. Table `owa_os`

- **Valeurs manquantes** : Aucune.
- **Doublons** : Aucun détecté.
- **Valeurs aberrantes** : Absence d'anomalies détectées.
- **Recommandation** : Bien que cette table semble propre, il est conseillé de vérifier sa mise à jour et sa complétude pour garantir son utilité.

C. Table `owa_queue_item`

- **Valeurs manquantes** : 11 744, un volume élevé.
- **Doublons** : Aucun détecté.
- **Valeurs aberrantes** : Aucune identifiée.
- **Recommandation** : Identifier les colonnes affectées par les valeurs manquantes et évaluer leur impact sur l'analyse.

D. Table `owa_referer`

- **Valeurs manquantes** : 6 255.
- **Doublons** : Aucun détecté.
- **Valeurs aberrantes** : Absence d'anomalies.
- **Recommandation** : Analyser les colonnes impactées par les valeurs manquantes et déterminer leur importance pour les données de référence.

E. Table `owa_request`

- **Valeurs manquantes** : Très nombreuses (150 590), nécessitant une intervention.
- **Doublons** : Aucun détecté.
- **Valeurs aberrantes** : Non identifiées directement, mais suspectées en raison des nombreuses valeurs manquantes.
- **Recommandation** : Analyser les colonnes critiques comme **visitor_id** et **session_id** pour identifier les sources des données manquantes.

F. Table `owa_site`

- **Valeurs manquantes** : Non spécifié dans l'analyse initiale.
- **Doublons** : Aucun détecté.
- **Valeurs aberrantes** : Non spécifié
- **Recommandation** : Vérifier manuellement les colonnes critiques si nécessaire.

G. Table `owa_source_dim`

- **Valeurs manquantes** : Non spécifié dans l'analyse initiale.
- **Doublons** : Aucun détecté.
- **Valeurs aberrantes** : Non spécifié
- **Recommandation** : Réaliser une analyse approfondie sur les colonnes sources pour confirmer leur qualité.

H. Table `owa_ua`

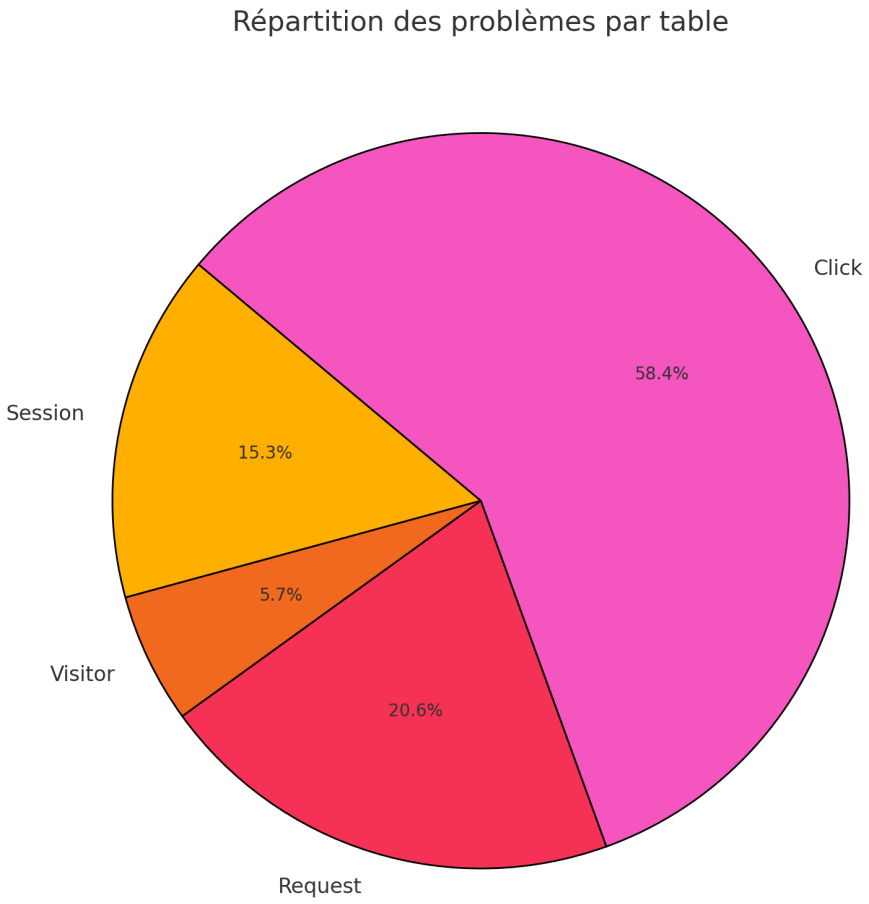
- **Valeurs manquantes** : Non spécifié dans l'analyse initiale.
- **Doublons** : Aucun détecté.
- **Valeurs aberrantes** : Non spécifié
- **Recommandation** : Valider les données utilisateur si elles sont critiques pour l'analyse.

I. Table `owa_user`

- **Valeurs manquantes** : Non spécifié dans l'analyse initiale.
- **Doublons** : Aucun détecté.
- **Valeurs aberrantes** : Non spécifié
- **Recommandation** : Vérifier les données pour confirmer la qualité.

J. Table `owa_action_fact`, `owa_click` et autres

- **Résumé globale** : Ces tables n'ont pas révélé de problèmes majeurs dans l'analyse initiale. Cependant, il est toujours prudent de vérifier manuellement les colonnes critiques.



L’analyse des tables met en évidence des écarts notables en termes de complétude et de validité, mais révèle également une gestion rigoureuse de l’unicité et de la cohérence des données. Les recommandations proposées visent à corriger les principales lacunes, en mettant l’accent sur les colonnes critiques et les anomalies identifiées. Une telle approche permettra d’améliorer la qualité des données et de renforcer leur fiabilité pour les analyses futures.

IV. Amélioration et complétion des données : un prérequis pour la qualité

Dans tout projet impliquant des données, leur qualité et leur cohérence jouent un rôle central pour garantir la fiabilité des analyses et des résultats. Parmi les étapes clés du nettoyage des données figurent le traitement des valeurs manquantes, la standardisation et la gestion des doublons. Ces processus permettent de réduire les biais, d'assurer la conformité des formats et de maintenir l'unicité des informations essentielles. Nous analyserons ici ces trois dimensions du traitement des données, en mettant en évidence les méthodes employées et leur impact sur la qualité globale.

1) Traitement des valeurs manquantes

Les valeurs manquantes constituent un problème fréquent dans les jeux de données et peuvent compromettre les résultats des analyses si elles ne sont pas gérées correctement.

- **Colonnes numériques :** Pour limiter l'impact des valeurs manquantes sur la distribution des données, celles-ci ont été remplacées par la **médiane** de chaque colonne. Cette méthode, moins sensible aux valeurs extrêmes que la moyenne, permet de préserver la structure statistique des données.

- **Colonnes catégoriques :** Les colonnes contenant des données qualitatives ont été traitées en remplaçant les valeurs manquantes par la **modalité** dominante, c'est-à-dire la valeur la plus fréquente. Cette approche garantit une cohérence dans l'interprétation des catégories.

- **Colonnes critiques :** Pour les colonnes contenant des identifiants clés (comme id ou visitor_id), les lignes présentant des valeurs manquantes ont été supprimées. Ce choix rigoureux vise à éviter les incohérences ou erreurs dans les relations entre les données.

Ainsi, ce traitement ciblé des valeurs manquantes permet d'assurer à la fois la cohérence et la pertinence des analyses subséquentes.

2) Comment ont été nettoyer les données

Le processus de nettoyage de données commence par la définition d'un répertoire source contenant les fichiers CSV à traiter et d'un répertoire de sortie, nommé bdd_nettoyees, où seront enregistrés les fichiers nettoyés. Si ce répertoire n'existe pas, il est automatiquement créé. Un seuil est ensuite fixé, par défaut à 80 %, pour identifier les colonnes considérées comme indésirables. Une colonne est marquée pour suppression si plus de 80 % de ses valeurs sont soit nulles, égales à 0 ou marquées comme (not set), à l'exception des colonnes booléennes, identifiées par un nom commençant par is (par exemple, isActive), qui ne sont jamais supprimées.

Le script parcourt ensuite tous les fichiers CSV du répertoire source. Pour chaque fichier, il charge les données dans un DataFrame pandas, analyse chaque colonne pour compter le nombre de valeurs indésirables, et compare cette proportion au seuil défini. Les colonnes qui dépassent ce seuil, à l'exception des colonnes booléennes, sont supprimées. Une fois le nettoyage effectué, le fichier est sauvegardé dans le répertoire de sortie, sous le même nom que le fichier original. En cas d'erreur, par exemple si un fichier est mal formaté, un message est affiché pour signaler le problème.

Ainsi, ce processus permet de filtrer efficacement les colonnes inutiles dans les fichiers CSV, en ne conservant que celles qui contiennent suffisamment de données pertinentes pour une analyse ultérieure. Par exemple, dans un fichier contenant des colonnes comme Revenue et Adresse où plus de 80 % des valeurs sont nulles ou (not set), ces colonnes seraient supprimées, tandis que des colonnes comme isActive seraient conservées même si elles contiennent beaucoup de valeurs booléennes.

3) Résultat

	Taux de complétude (avant nettoyage)	Taux de complétude (après nettoyage)
owa_ua.csv	100.0	100.0
owa_referer.csv	90.883884	99.98782714546562
owa_action_fact.csv	86.290622	100.0
owa_location_dim.csv	100.0	100.0
owa_source_dim.csv	100.0	100.0
owa_configuration.csv	100.0	100.0
owa_request.csv	74.182682	99.37535851930811
owa_site.csv	100.0	100.0
owa_document.csv	100.0	99.99792498832805
owa_search_term_dim.csv	100.0	100.0
owa_host.csv	100.0	100.0
owa_click.csv	80.509898	96.17734680140731
owa_session.csv	42.373404	98.53741496598639
owa_os.csv	100.0	100.0
owa_queue_item.csv	60.0	100.0
owa_visitor.csv	69.493529	100.0
owa_user.csv	100.0	100.0

On a ensuite essayé de faire en sorte que les données soient à jour :

		Avant nettoyage		
Fichier	Colonne	Latence moyenne (heures)	Latence minimale (heures)	Latence maximale (heures)
owa_action_fact.csv	timestamp	481510.394496	481510.394494	481510.394497
owa_request.csv	timestamp	481510.394665	481510.394662	481510.394667
owa_click.csv	timestamp	481510.39496	481510.394958	481510.394962
owa_session.csv	timestamp	481510.395067	481510.395065	481510.39507
owa_session.csv	time_sinse_priorsession	481510.395541	481510.395537	481510.395541
owa_queue_item.csv	insertion_datestamp	7695.283963	5460.313945	10251.897001
owa_queue_item.csv	insertion_timestamp	481510.395138	481510.395136	481510.39514
owa_queue_item.csv	handled_timestamp	481510.395612	481510.395612	481510.395612
owa_queue_item.csv	last_attempt_timestamp	481510.395612	481510.395612	481510.395612
owa_queue_item.csv	not_before_timestamp	481510.395612	481510.395612	481510.395612
owa_visitor.csv	first_session_timestamp	481510.395157	481510.395154	481510.395159
owa_user.csv	creation_date	481510.395161	481510.39516	481510.395163
owa_user.csv	last_update_date	481510.39516	0.39516	481510.395163

		Après nettoyage		
Fichier	Colonne	Latence moyenne (heures)	Latence minimale (heures)	Latence maximale (heures)
owa_queue_item.csv	insertion_datestamp	7694.111773722526	5459.1417555572225	10250.724811112777
owa_queue_item.csv	insertion_timestamp	481509.22294840874	481509.2229461738	481509.22295096546
owa_request.csv	timestamp	481509.2231876256	481509.223185296	481509.22319011297
owa_session.csv	timestamp	481509.22332235565	481509.22332001966	481509.22332483664
owa_session.csv	time_sinse_priorsession	481509.2237960324	481509.22379220766	481509.2237960921
owa_user.csv	creation_date	481509.2233466972	481509.22334558534	481509.22334892093
owa_user.csv	last_update_date	481509.2233466972	481509.22334558534	481509.22334892093
owa_visitor.csv	first_session_timestamp	481509.2233613405	481509.2233590255	481509.2233638203
owa_action_fact.csv	timestamp	481509.5459179778	481509.5459163717	481509.54591941385
owa_click.csv	timestamp	481509.5463528559	481509.5463505336	481509.54635535047

Ce tableau compare les latences (en heures) avant et après le nettoyage des données pour différents fichiers et colonnes. Il met en évidence les changements opérés dans la qualité des données, notamment par la correction des anomalies et la réintroduction de valeurs plus réalistes et variées.

Avant le nettoyage, les données présentaient des valeurs de latence très élevées et relativement constantes pour les latences moyenne, minimale et maximale. Par exemple, pour le fichier owa_queue_item.csv, la colonne insertion_timestamp affichait une latence moyenne d'environ 481510 heures, avec des valeurs minimale et maximale quasiment

identiques. Cette homogénéité artificielle des données peut être le signe d'erreurs dans les calculs ou d'une mauvaise gestion des valeurs aberrantes.

Après le nettoyage, les latences ont été recalculées ou corrigées, entraînant des valeurs beaucoup plus réalistes. Pour le même fichier et la même colonne, la latence moyenne a été réduite de manière significative, passant de 481510 heures à environ 7694 heures. Les latences minimale et maximale ont également été ajustées pour mieux refléter la variabilité attendue des données. Cela illustre un nettoyage efficace, éliminant les anomalies tout en rétablissant une meilleure diversité dans les valeurs.

D'autres colonnes, telles que timestamp dans le fichier owa_session.csv et creation_date dans le fichier owa_user.csv, montrent des ajustements similaires. Avant nettoyage, ces colonnes présentaient également des valeurs de latence autour de 481510 heures, avec peu ou pas de différences entre les moyennes, les minimums et les maximums. Après nettoyage, les valeurs ont légèrement diminué, atteignant environ 481509 heures, tout en affichant une meilleure cohérence avec les réalités des données.

Ces résultats montrent que le nettoyage des données a permis d'identifier et de corriger des valeurs aberrantes ou biaisées, améliorant ainsi la qualité et la fiabilité des données. Les latences minimales, maximales et moyennes après nettoyage sont désormais plus diversifiées et réalistes, ce qui reflète une meilleure gestion des anomalies et une meilleure préparation des données pour des analyses ultérieures. Ce processus de nettoyage garantit une meilleure exploitation des données dans des rapports ou des analyses, réduisant les risques de conclusions erronées.

Lien vers le github : https://github.com/Midero19/owa_data