Data Management M2

SPRINT 3

Management de Projets (Scrum)

ÉQUIPE

Nadir BOUDJERIDA
Scrum Master

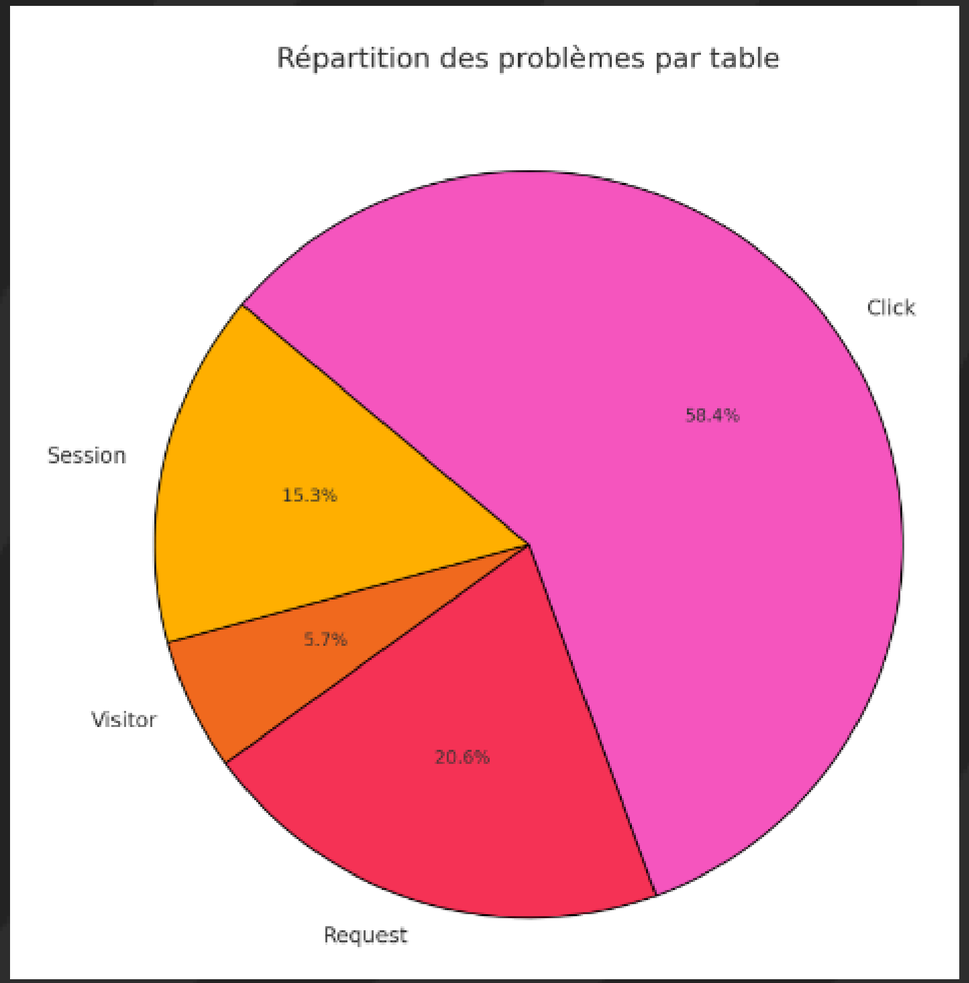
Jean-Louis HU
Data Analyst

Thomas DAHROUJ
Product Owner

ANALYSE PRÉLIMINAIRES DES DONNÉES

Critères de qualité de données
Doit catégoriser les actions en groupes valides (exemple : 'publish', 'edit'). Temporalité : La liste des groupes doit être mise à jour régulièrement pour inclure les nouvelles catégories ou actions. Traçabilité : Chaque groupe doit être documenté avec sa date de création, sa version et ses mises à jour. Confidentialité : Si certaines actions sont sensibles, elles doivent être protégées et ne pas apparaître en clair.
Doit fournir un contexte descriptif de l'action (exemple : 'dataset uploaded by user'). Doit avoir une longueur maximale de 100 caractères. Ne doit pas contenir de données sensibles (ex. : noms d'utilisateur, adresses email).
Doit être valide et correspondre à une action définie (exemple : 'create', 'modify'). Doit respecter un format spécifique (ex. : lettres minuscules et underscores).
Doit correspondre à une publicité valide unique si utilisé
Doit référencer une campagne existante et unique
Doit être alphanumérique et non vide si utilisé
Doit être cohérent avec cv1_name et non nul si utilisé
Doit être alphanumérique et non vide si utilisé
Doit être cohérent avec cv2_name et non nul si utilisé
Doit être alphanumérique et non vide si utilisé
Doit être cohérent avec cv3_name et non nul si utilisé
Doit être alphanumérique et non vide si utilisé
Doit être cohérent avec cv4_name et non nul si utilisé
Doit être alphanumérique et non vide si utilisé
Doit être cohérent avec cv5_name et non nul si utilisé
Doit être entre 1 et 31. Cohérence : Le jour doit être valide par rapport au mois et à l'année indiqués (exemple : 30 février est invalide).
Doit être entre 0 (Dimanche) et 6 (Samedi). Comparabilité : Les codes doivent correspondre aux normes internationales (exemple : ISO 8601).
Doit être entre 1 et 366. Temporalité : L'année doit être vérifiée pour savoir si elle est bissextile.
Doit être un entier positif ou nul. Temporalité : Les calculs doivent être actualisés pour chaque nouvelle session. Complétude : Ces champs ne doivent pas être manquants si des sessions sont enregistrées.
Doit être un entier positif ou nul. Temporalité : Les calculs doivent être actualisés pour chaque nouvelle session. Complétude : Ces champs ne doivent pas être manquants si des sessions sont enregistrées.
Doit référencer un document existant et être numérique et unique
Doit correspondre à un hôte valide et unique
Unicité : L'id doit être unique pour chaque enregistrement. Non-nullité : L'id ne doit pas être nul. Conformité au type de données : L'id doit être un entier de type int64. Valeur correcte : L'id doit être attribué de manière cohérente, généralement de manière croissante ou séquentielle. Pas de valeurs erronées : L'id ne doit pas être négatif si le système n'accepte que des identifiants positifs.
Doit être dans un format IPv4 ou IPv6 valide. Confidentialité : Les adresses IP doivent être anonymisées si elles ne sont pas nécessaires pour l'analyse.
Booléen (0 ou 1)

CONTRÔLER DES QUALITÉS DES DONNÉES



INDICATEURS CLÉS

	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

		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	481510.395161	481510.39516	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

AMÉLIORATION ET COMPLÉTION DES DONNÉES

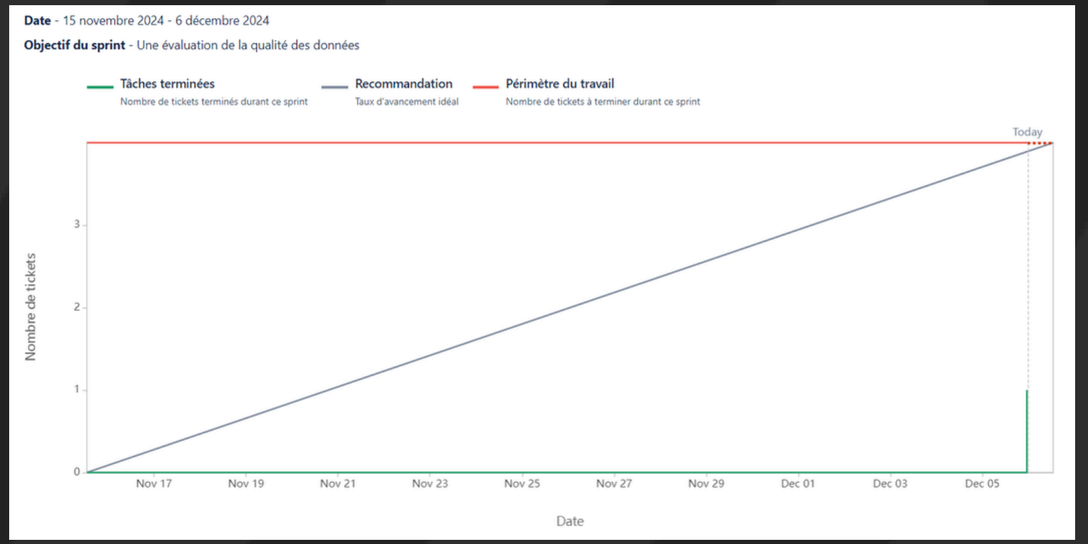
- Traitement des valeurs manquantes
- Nettoyer les données
- Analyse des différences avant/après nettoyage
- Conclusion

NOS RÉALISATIONS PASSÉES

- Correction de la description
- Définition des critères de qualités de données pour chaque tables
- Vérification des anomalies dans la base de données
- Nettoyage des données en fonctions des anomalies

AVANCEMENT DES TÂCHES

Résumé	État	Sprint	Personne assignée	Date d...
Effectuez une analyse initiale des données pour identifier les ...	TERMINÉ(E)	Tableau Sprint 3	BN Boudjerida Nadir	22 nov. 2024
Définir des indicateurs de performances (KPI) pour surveiller ...	TERMINÉ(E)	Tableau Sprint 3	JH Jean-Louis HU	23 nov. 2024
Contrôler la qualité des données (tests)	TERMINÉ(E)	Tableau Sprint 3	TD Thomas DAHROUJ	29 nov. 2024
Ajoutez des données manquantes ou améliorez la qualité de...	TERMINÉ(E)	Tableau Sprint 3	JH Jean-Louis HU	30 nov. 2024



NOS RÉALISATIONS FUTURES

Essayer d’automatiser les mises à jour des données

Mettre en place une solution de gestion de la qualité des données sur la plateforme OpenMetadata.

Analyser l'avancement et finaliser le projet à l'aide d'un Burndown chart.

Proposer une présentation d’un projet en s’appuyant sur un article et un clip vidéo, tout en élaborant des recommandations stratégiques en matière de gouvernance des données.

LES DIFFICULTÉS RENCONTRÉES

- Analyse des données
- Nettoyage des données
- Correction des champs des tables

Base de Données :

THOMAS DAHROUJ - JEAN-LOUIS HU - NADIR BOUDJERIDA