CulturaPass – Livrables du partiel (Bloc 1)

Source principale : Énoncé de l'étude de cas fournie (CSV exemples et contexte). Références réglementaires : RGPD (art. 5 et 6), recommandations CNIL sur la durée de conservation et l'usage des données.

1. Reformulation du besoin métier

CulturaPass souhaite structurer et fiabiliser des fichiers CSV hétérogènes transmis par des partenaires afin de créer une base de données exploitable pour des analyses statistiques et, à terme, des modèles d'IA, tout en respectant les exigences du RGPD.

2. Cartographie des données

Les fichiers identifiés sont : evenements.csv, inscriptions.csv et organisateurs.csv. Les relations principales sont : un événement possède plusieurs inscriptions, et un organisateur peut proposer plusieurs événements.

3. Organisation et stockage

Le stockage recommandé repose sur une base relationnelle centralisée (PostgreSQL). Une zone de staging conserve les CSV bruts pour traçabilité. Les tables cibles sont normalisées et reliées par des clés primaires et étrangères.

4. Nettoyage et harmonisation des données

Les dates sont converties au format ISO (YYYY-MM-DD). Les catégories et villes sont harmonisées via des tables de correspondance. Les valeurs manquantes sont conservées comme NULL et, si nécessaire, imputées uniquement pour les analyses avec un indicateur explicite d'imputation.

5. Modèle de données (synthèse)

Les tables principales sont : events, registrations, organizers, cities et categories. Les inscriptions constituent la table de faits principale pour les analyses de fréquentation.

6. Analyses rendues possibles

La base permet de calculer le taux de participation par événement, d'identifier les événements sous-utilisés et de détecter des pics de fréquentation par période.

7. Tests de performance

Des index sur les clés de jointure (event_id, city_id, category_id) améliorent les performances. Pour de gros volumes, un partitionnement temporel est recommandé.

8. Enjeux réglementaires et éthiques

Le traitement respecte le principe de minimisation et de limitation de la durée de conservation. Les données utilisateurs doivent être pseudonymisées. Une information claire des personnes et, si nécessaire, une analyse d'impact (DPIA) sont requises.

9. Limites des données

La présence de valeurs manquantes et l’hétérogénéité des formats limitent la précision des analyses. L’absence d’identifiant utilisateur stable empêche un suivi longitudinal fiable.

10. Conclusion

La structuration proposée permet une exploitation analytique fiable des données CulturaPass. La priorité doit être donnée à la qualité des données, à la traçabilité et à la conformité réglementaire avant tout usage avancé en intelligence artificielle.