Étude de cas — CulturaPass

Structuration, gouvernance et exploitation des données culturelles

Livrable académique : Mastère RNCP Niveau 7 — Bloc 1

Auteur : Candidat
Date : 24 janvier 2026

Table des matières

Résumé exécutif

Ce document est un livrable académique de niveau Mastère RNCP (Niveau 7) préparé pour le Bloc 1 — « Analyser et structurer le besoin client visant le développement d'une solution d'Intelligence Artificielle ». Il reprend et développe l'étude de cas fournie pour CulturaPass afin de répondre aux attentes du référentiel RNCP : analyse du besoin, audit et cartographie des données, proposition de solutions techniques,gestion de la conformité RGPD et prise en compte des enjeux éthiques et d'inclusion. Le référentiel d'évaluation précise que l'étude attend un rapport conséquent (≈30 pages) et une couverture détaillée des compétences C1.1 à C1.9. ■filecite■turn1file0■

1. Contexte et objectifs

CulturaPass agrège des données d'événements culturels et d'inscription provenant de partenaires sous forme de fichiers CSV hétérogènes. L'objectif principal est d'industrialiser l'ingestion, le nettoyage, la structuration et la gouvernance des données pour produire des indicateurs fiables et préparer l'intégration d'algorithmes prédictifs (prévision de fréquentation, segmentation des publics). Les contraintes comprennent la diversité des formats, des valeurs manquantes, la nécessité de traçabilité et la conformité RGPD. Le jeu d'exemples fourni avec l'énoncé sert de base pour le profilage initial. ■filecite■turn0file0■

2. Reformulation du besoin métier (compétence C1.1)

Reformulation claire et synthétique du besoin : CulturaPass souhaite disposer d'une plateforme de données fiable et traçable permettant d'extraire des indicateurs de fréquentation, d'optimiser l'allocation de capacité événementielle, et d'alimenter des modèles de prédiction, tout en respectant les obligations légales et l'éthique. Le périmètre initial comprend les fichiers : evenements.csv, inscriptions.csv, organisateurs.csv. ■filecite■turn0file0■
Hypothèses opérationnelles : - Les partenaires fournissent des exports CSV (ponctuels ou récurrents). - Les identifiants peuvent être pseudonymes ; les données personnelles doivent être traitées selon RGPD. - Volumétrie attendue : de l'ordre de quelques milliers à quelques centaines de milliers de lignes par an selon l'échelle des partenaires (voir section 4).

3. Cartographie et audit des données (compétence C1.2)

Objectif : dresser un inventaire exhaustif des sources, typologies et formats et évaluer la faisabilité technique.

3.1. Sources identifiées

• evenements.csv — métadonnées événement (event_id, titre, catégorie, ville, date de début/fin, capacité)
• inscriptions.csv — enregistrements des inscriptions (registration_id, event_id, user_id/pseudonyme, date_inscription, statut, âge)
• organisateurs.csv — informations organisateur (organizer_id, nom, contact, ville)
• Méta-informations : logs d'ingestion, fichiers bruts conservés en zone de staging

3.2. Profiling initial (exemples observés)

Extraits du jeu d'exemples indiquent : champs vides pour max_capacity et user_age, formats de date hétérogènes, doublons partiels sur les inscriptions. Ces constats motivent des règles d'imputation documentées et des contrôles de qualité automatiques. ■filecite■turn0file0■

4. Volumétrie et typologie des données (C1.2, C1.6)

Pour dimensionner la solution, il est nécessaire d'estimer la volumétrie. Trois scénarios sont proposés :

Scénario	Estimation volumétrique / préconisation
Petit (pilot)	≈ 10k enregistrements/an, <100 Mo, base PostgreSQL mono-instance
Moyen (déploiement régional)	≈ 200k-1M enregistrements/an, 0.5–3 Go, PostgreSQL avec indexation et vues
Large (national)	≥ 10M enregistrements/an, plusieurs dizaines de gigaoctets, partitionnement te

Choix par défaut pour un projet académique : PostgreSQL (relationnel) avec zone de staging fichiers (S3 ou équivalent) et stockage des CSV bruts pour auditabilité. Ce choix répond aux besoins d'intégrité référentielle et de requêtage analytique (C1.6).

5. Solutions techniques proposées (compétences C1.4, C1.6)

5.1. Architecture cible (haut niveau)

• Zone de réception (partenaires) → stockage brut (S3 / storage)
• Pipeline ETL (ex : Airflow / Prefect) → staging tables (stg_events, stg_registrations, stg_organizers)
• Transformation (dbt ou scripts) → modèle dimensionnel (events, registrations, organizers, cities, categories, date_dim)
• Datamart / vues matérialisées pour reporting et export pour modèles IA
• Système de gouvernance : catalogage (Metacat), logs d'ingestion, gestion des accès (RBAC)

5.2. Choix de la base : SQL vs NoSQL

Argumentaire : Les données sont fortement relationnelles (événements ↔ inscriptions ↔ organisateurs). Un SGBD relationnel (PostgreSQL) est recommandé. Un stockage NoSQL (ex : document store) n'apporterait pas d'avantage majeur ici sauf pour logs non-structurés ou très forte scalabilité en écriture. Le référentiel RNCP demande de justifier le choix en fonction de la volumétrie et de la typologie (C1.6). ■filecite■turn1file0■

6. Règles de traitement, nettoyage et intégration (C1.8, C1.9)

6.1. Principes généraux

• Traçabilité : conserver fichier brut + métadonnées (source_file, ingest_ts, row_hash)
• Idempotence : transformations répétables sans effets indésirables
• Conservatisme pour imputations : ne pas modifier les données d'origine ; documenter toute imputation
• Validation forte des clés primaires et des contraintes référentielles au chargement

6.2. Règles concrètes (exemples)

• Dates → format ISO (YYYY-MM-DD). Si invalide → flag sur la ligne et mise en quarantaine.
• Capacité (max_capacity) manquante → marquer NULL ; pour analyses explorer estimation par catégorie/ville mais ne pas remplacer définitivement.
• Doublons d'inscription : key = (event_id, user_id/pseudonyme, registration_date) ; garder la plus récente si conflit entre partenaires.
• Imputation d'âge : uniquement pour analyses agrégées ; conserver un marqueur age_imputed=true.

6.3. Exemple de pipeline ETL (pseudocode)

```
# Pseudocode ETL (high-level) 1. Ingest file -> store in raw_bucket with metadata (filename, timestamp, partner_id) 2. Load raw CSV into staging table stg_xxx as TEXT (no parsing) 3. Profiling: compute missing%, unique counts, sample values 4. Apply parsing rules -> normalized staging table (dates, trims, lower/upper) 5. Map categories and cities via lookup tables; unresolved flagged for manuel review 6. Deduplicate according to rules -> insert/update into production tables in transaction 7. Log ingest result into ingest_log (rows_in, rows_out, errors)
```

7. Contraintes réglementaires et éthiques (compétence C1.5)

Points essentiels : base légale, durée de conservation, droits des personnes, pseudonymisation/anonymisation, DPIA si profiling à large échelle. Le référentiel souligne l'importance d'intégrer la dimension éthique et réglementaire dans la proposition technique. ■filecite■turn1file0■

7.1. Actions requises

• Rédiger la finalité par traitement et définir la base légale (consentement / intérêt légitime / exécution contractuelle)
• Mettre en place des durées de conservation automatisées et des routines de purge/anonymisation
• Pseudonymiser les identifiants utilisateurs dans les datamarts et anonymiser les jeux rendus publics
• Réaliser une DPIA si utilisation d’algorithmes de profilage susceptibles d’impacter les droits

8. Accessibilité et inclusion (prise en compte des situations de handicap)

Le référentiel RNCP insiste sur l'adaptation de la communication et des livrables aux personnes en situation de handicap (A1.1, C4.3). Livrables attendus : versions accessibles des dashboards (contraste, textes alternatifs), transcriptions et supports imprimés compatibles, sessions de restitution adaptées.

9. Correspondance avec le référentiel d’évaluation — Bloc 1

Cette section explicite comment le présent livrable couvre les compétences C1.1 à C1.9 et les critères d’évaluation du référentiel RNCP.
• C1.1 — Analyse du besoin, identification des sources — couvert en sections 1–3; sources listées et reformulation réalisée. ■filecite■turn0file0■
• C1.2 — Audit et cartographie (volumétrie, typologie) — sections 3–4; scénarios de volumétrie fournis.
• C1.4 — Proposition de solutions techniques — section 5 architecture et justification choix SGBD.
• C1.5 — Dimension éthique et réglementaire — section 7 recommandations et actions (DPIA, pseudonymisation). ■filecite■turn1file0■
• C1.6 — Choix du gestionnaire de base et justification selon volumétrie — section 4–5.
• C1.8/C1.9 — Nettoyage, imputation et intégration — section 6 avec pseudocode ETL.

10. Plan d’implémentation et rétroplanning (compétence C4.1)

Phases projet et livrables attendus (chronologie indicative sur 8 semaines)

Période	Activités
Semaine 1	Kickoff, collecte fichiers partenaires, mise en place zone de staging
Semaine 2-3	Profiling approfondi, dictionnaires de mapping (villes, catégories)

Semaine 4	Développement pipelines ETL & tests unitaires
Semaine 5	Chargement initial en base, QA et correction
Semaine 6	Vues analytiques, dashboards QA, tests de performance
Semaine 7	Documentation RGPD, DPIA (préliminaire), livrables pour jury
Semaine 8	Préparation soutenance : rapport final, notebook, annexes

11. Jeux d’essais, tests et indicateurs qualité (C1.2, C1.10)

KPIs qualité de données proposés :

• Taux de valeurs manquantes par champ
• Taux de doublons détectés et corrigés
• Nombre de lignes en quarantaine (format/date invalides)
• Temps moyen d’ingestion par fichier
• Couverture des imputations avec indicateurs (ex: age_imputed %)

12. Annexes : exemples techniques

12.1. Extrait DDL (PostgreSQL)

```
CREATE TABLE cities (city_id SERIAL PRIMARY KEY, city_name TEXT UNIQUE); CREATE TABLE
categories (category_id SERIAL PRIMARY KEY, category_label TEXT UNIQUE); CREATE TABLE
organizers (organizer_id TEXT PRIMARY KEY, city_id INT REFERENCES cities(city_id),
contact JSONB); CREATE TABLE events (event_id TEXT PRIMARY KEY, organizer_id TEXT
REFERENCES organizers(organizer_id), city_id INT REFERENCES cities(city_id),
category_id INT REFERENCES categories(category_id), start_date DATE, end_date DATE,
max_capacity INT, source_file TEXT, ingest_ts TIMESTAMP DEFAULT now()); CREATE TABLE
registrations (registration_id TEXT PRIMARY KEY, event_id TEXT REFERENCES
events(event_id), user_id TEXT, user_age INT, age_imputed BOOLEAN DEFAULT FALSE,
registration_date DATE, attendance_status TEXT, source_file TEXT, ingest_ts TIMESTAMP
DEFAULT now());
```

12.2. Exemple de requête : taux de participation

```
SELECT e.event_id, e.start_date, COUNT(r.registration_id) AS registrations,
e.max_capacity, CASE WHEN e.max_capacity IS NOT NULL THEN ROUND(100.0 *
COUNT(r.registration_id) / e.max_capacity, 1) ELSE NULL END AS pct_capacity FROM events
e LEFT JOIN registrations r ON r.event_id = e.event_id GROUP BY e.event_id,
e.start_date, e.max_capacity;
```

12.3. Bibliographie et sources

• Cahier d’exemples et énoncé fourni pour l’étude de cas (CulturaPass). ■filecite■turn0file0■
• Référentiel « Chef de projet data & intelligence artificielle » (RNCP) — modalité et critères d’évaluation Bloc 1. ■filecite■turn1file0■

Annexe : Liste de contrôle pour la soutenance (Bloc 1)

• Le rapport couvre C1.1 à C1.9 avec preuves documentées.
• Inventaire des sources et profiling effectué.
• Architecture technique proposée avec justification.
• Plan RGPD et actions (pseudonymisation, DPIA déclarée si nécessaire).
• Deliverables techniques prêts : DDL, scripts ETL, requêtes exemples.
• Supports adaptés pour personnes en situation de handicap.

Conclusion

Ce livrable est conçu pour répondre aux attentes d’un rapport académique de niveau Mastère RNCP (Bloc 1). Il documente la compréhension du besoin, l'audit des données, les solutions techniques et les aspects réglementaires et éthiques. Il propose également un plan d’action et des annexes techniques exploitables pour une mise en œuvre opérationnelle. Pour la soutenance, il est recommandé d’accompagner ce rapport par : un notebook reproductible, des extraits de pipelines ETL et un jeu d’essais démontrant la qualité des données.

Fait le : 24 janvier 2026