Projet de Base de Données

Arthur Hauguel - Loïc Mémeteau

---

1. Contexte du projet

L'objectif du projet est de construire une base de données relationnelle pour Music&Films Company.

Le travail s'est fait en deux temps :

1. Partir d'un schéma imparfait avec des données manquantes ou erronées, le but étant d'identifier les problèmes et comment les régler.
2. Nettoyer ces données et améliorer le schéma pour obtenir une base plus cohérente, mieux normalisée et plus résistante.

---

2. Schéma initial : description rapide

Le schéma de départ est composé des tables suivantes :

- ARTISTE(idArtiste, nom, pays, naissance, type)
  Artistes avec un nom, un pays en clair (ex : "France", "FR", vide…), une date de naissance, et un type (R, A, M sans contrainte).

- FILM(idFilm, titre, annee, genre, idRealisateur, codePays)
  Films avec titre, année, genre, un réalisateur (clé étrangère vers ARTISTE), et un code pays texte libre.

- ALBUM(idAlbum, titre, annee, genre, idArtistePrincipal, label)
  Albums avec un artiste principal mais sans clé étrangère vers ARTISTE.

- UTILISATEUR(idUser, nom, age, pays)
  Utilisateurs avec nom, âge et pays, sans contraintes particulières.

- AVIS(idAvis, idUser, idFilm, idAlbum, note, commentaire, dateAvis)
  Une seule table pour les avis sur films et albums, sans aucune clé étrangère, et une note entière non bornée.

- PARTICIPE_FILM(idArtiste, idFilm, role)
  Table de relation artiste–film, sans clé primaire ni clé étrangère.

Chaque table (sauf PARTICIPE_FILM) a une clé primaire, et les dépendances internes sont faciles à implémenter. En revanche, la base manque clairement de contraintes sur certaines variables.

---

3. Problèmes observés dans les données

3.1 Artistes, films et albums

- Dans ARTISTE :
    - pays vides ou incohérents,
    - type manquant ou erroné (type vide, NULL…),

- Dans FILM :
    - titres vides,
    - genres vides ou incohérents,
    - réalisateurs inexistants si les contraintes ne sont pas bien appliquées,
    - codes pays hétérogènes (FR, France, USA, vide, NULL).

- Dans ALBUM :
    - artiste principal qui n'existe pas dans ARTISTE,
    - albums sans titre ni année,
    - labels manquants.

3.2 Utilisateurs et avis

- UTILISATEUR :
    - utilisateurs sans nom, sans âge, sans pays,
    - aucun contrôle sur l'âge.

- AVIS :
    - avis rattachés à des films ou albums inexistants,
    - notes hors échelle (-2, 19, etc.),
    - avis avec utilisateur inexistant,
    - champs essentiels vides (date, commentaire, idFilm/idAlbum…).

3.3 Participation aux films

Dans PARTICIPE_FILM :

- lignes pointant sur des artistes ou des films inexistants,
- rôle NULL ou incohérent,
- possibilité de doublons (même artiste, même film, même rôle) faute de clé primaire.

De plus, tous les id sont aléatoires, ils commencent à 10, 100 voir même 1000 alors que nous n'avons même pas plusieurs dizaines de lignes.

Même si on est proche de la 3NF au niveau de chaque table séparément, la qualité globale de la base est mauvaise : données incohérentes, références absentes, valeurs aberrantes et une faible résistance à de possibles nouvelles données.

---

4. Nettoyage des données

Avant de toucher au schéma, une première phase a consisté à corriger/supprimer les données problématiques.

4.1 ALBUM

- Suppression des albums sans titre (titre IS NULL).
- Complétion des données manquantes lorsqu'une valeur fiable est connue (mise à jour du label, ajout de l'année pour certains albums.)
- Suppression des albums dont l'année reste inconnue et non récupérable.

4.2 ARTISTE

- Suppression d'un artiste incohérent, puis réinsertion avec des informations complètes.
- Complétion des pays (pays = 'FR' pour certains),
- Ajout ou correction de dates de naissance,
- Correction du type (M, R, A),
- Ajout d'artistes manquants (ex. Melville, Polanski) pour cohérer avec les films existants.

4.3 AVIS

- Suppression d'un avis inutilisable.
- Rebornage automatique des notes dans [0 ; 5].
- Fixation d'une date d'avis cohérente pour certaines lignes.

4.4 FILM, PARTICIPE_FILM et UTILISATEUR

- FILM :
    - suppression d'un film inutilisable,
    - correction de titres, genres et réalisateurs,
    - harmonisation de certains pays.
- PARTICIPE_FILM :
    - correction de rôles incohérents,
    - suppression de participations vers un artiste inexistant.
- UTILISATEUR :
    - suppression des utilisateurs sans nom.

Cette phase de nettoyage est nécessaire pour réduire adapter les données déjà présentes avec d'ajouter des contraintes.

---

5. Amélioration et normalisation du schéma

Une fois les données nettoyées, le schéma a été transformé pour mieux respecter les principes de normalisation. Techniquement, cela passe par la création de tables *COPIE*, le transfert des données, puis le renommage. Cette étape s'est rapidement avérée essentielle pour l'ajout des clés étrangères manquantes et de contraintes

L'option "CASCADE" a été ajoutée à ON UPDATE et ON DELETE pour permettre une cohérence durable entre les tables liées lors de futures modifications.

5.1 ARTISTE, pays et métier

L'ancienne table ARTISTE mélangeait des informations de nature différente. Elle a été éclatée en trois tables :

- ARTISTE(idArtiste, nom, naissance)
  Nom obligatoire, date de naissance éventuellement inconnue.

- PAYS_ARTISTE(idArtiste, pays_iso2)
  Clé primaire (idArtiste, pays_iso2).
  Permet de gérer un artiste avec zéro, un ou plusieurs pays. Le code ISO2 peut être NULL si le pays est inconnu.

- METIER_ARTISTE(idArtiste, metier)
  Un métier principal par artiste (R, A, M), avec clé primaire sur idArtiste.

Ces tables sont liées par des clés étrangères avec ON UPDATE CASCADE et ON DELETE CASCADE.

5.2 UTILISATEUR et pays

Même logique pour utilisateurs :

- UTILISATEUR(idUser, nom, age)
  Nom et âge NOT NULL, avec une contrainte CHECK(age < 125) pour éliminer les valeurs aberrantes.

- PAYS_UTILISATEUR(idUser, pays_iso2)
  Clé primaire (idUser, pays_iso2), pays en code ISO2 obligatoire cette fois-ci (contrairement aux artistes).

5.3 FILM, PAYS_FILM et PARTICIPE_FILM

Pour les films :

- FILM(idFilm, titre, annee, genre, idRealisateur)
  Titre, genre et réalisateur sont obligatoires. L'année peut rester NULL si inconnue.

- PAYS_FILM(idFilm, pays_iso2)
  Clé primaire (idFilm, pays_iso2).
  Permet de gérer les films coproduits par plusieurs pays. Le pays peut être NULL si l'information manque.

- PARTICIPE_FILM(idArtiste, idFilm, role)
  Clé primaire (idArtiste, idFilm), références vers ARTISTE et FILM.
  On évite ainsi les doublons et on garantit que chaque participation pointe vers un artiste et un film existants.

5.4 ALBUM et séparation des avis

Côté albums :

- ALBUM(idAlbum, titre, annee, genre, idArtistePrincipal, label)
  Titre et genre obligatoires, artiste principal obligatoire (clé étrangère vers ARTISTE), année et label facultatifs.

Pour les avis, la grosse amélioration est la séparation :

- AVIS_FILM(idAvisFilm, idUser, idFilm, note, commentaire, dateAvis)
- AVIS_ALBUM(idAvisAlbum, idUser, idAlbum, note, commentaire, dateAvis)

Dans les deux cas :

- idUser et idFilm/idAlbum sont des clés étrangères,
- la note est bornée par CHECK(note BETWEEN 0 AND 5),
- la date d'avis est obligatoirement renseignée,
- le commentaire reste facultatif.

L'ancienne table AVIS était ambiguë, maintenant on sait clairement si un avis porte sur un film ou sur un album.

5.5 Réindexation des identifiants

Enfin, une étape finale réajuste certains identifiants (albums, avis, films, utilisateurs) en les décalant pour obtenir des plages plus propres (à partir de 0 ou proche). Les contraintes CASCADE garantissent que ces modifications ne cassent pas les liens entre les tables.

---

6. Bilan

- Données nettoyées (suppression/correction des lignes irrécupérables) ;
- Séparation du schéma originel en tables plus cohérentes (PAYS_*, METIER_ARTISTE, AVIS_*)
- Contraintes d'intégrité rajoutées (clés étrangères systématiques avec ON UPDATE/ON DELETE, CHECK, NOT NULL)
- Gestion plus résistante des possibles cas extrêmes (gestion des multiples nationalités/métiers, avis distincts pour films et albums) et aux futurs modifications

(normalisation des numeros d’ids et ajout de l’option CASCADE pour les clés étrangères.

La base est maintenant plus lisible et prête à être utilisée pour des requêtes de statistiques, de recommandations, d’analyses, ou encore l’ajout de nouvelles données sans être polluée par des incohérences.