BDLE 2020-2021

Projet n°1

SQL à large échelle,

Application aux données de IMDB

Tâche 1

A partir du notebook du TP1 et des vues donnant des informations détaillées sur les œuvres (TitleDetail) et les personnes (NameDetail) :

TitleDetail(Id, title, production_year, kind, property, value)

NameDetail (Id, name, gender, property, value)

On remarque que l'attribut *property* décrit de nombreuses propriétés. Par exemple une œuvre peut avoir les propriétés : budget, countries, genres, languages, locations, filming_dates, *etc*. Une personne peut avoir les propriétés : birth date, interviews, article, salary history, *etc.*

Restructurer la base de manière à définir des relations plus spécifiques. Définir au moins 4 relations concernant les personnes et 4 relations concernant les œuvres. Ne pas vous limiter aux relations proposées ci-dessous qui sont données à titre d'exemple. Votre réponse doit présenter au moins 4 propriétés qui ne sont pas mentionnées ci-dessous.

Schéma des œuvres :

Langue (titleID, title, langue) la langue d'une œuvre, plusieurs tuples si plusieurs langues.

Genre (titleID, title, genre) le genre d'une œuvre

Budget (titleID, title, year, amount, currency)

Location (titleID, title, location, country)

ReleasedCountries (titleID, title, country)

Admissions (titleID, title, country, date, nb_entries) le nombre d'entrées au cinéma dans un pays pour un film à une date.

Schéma des personnes :

Person (personID, name, birthdate, deathdate)

Citizenship (personID, name, country)

Article (personID, name, magazine, year, country) un magazine dans un pays a publié à une date *year* un article sur une personne.

Salary (personID, name, titleID, title, year, amount, currency)

exple de tuple de Salary: Di Caprio a gagné 59M USD dans Inception en 2010

Tâche 2

L'objectif est d'analyser les données selon plusieurs dimensions.

Proposer un schéma en étoile où un fait est une association entre une personne, une œuvre et un rôle. Proposer des dimensions sur les personnes et les œuvres. Proposer au moins une dimension hiérarchique ayant 3 niveaux.

Tâche 3

Enrichir des informations de IMBD avec d'autres données, par exemple venant de wikidata, dbpedia.

Ajouter des propriétés pour une personne, une œuvre ou un lieu.

Tâche 4

Essayer de traiter la totalité des données de IMDB (pas seulement l'extrait sample01), et décrire très brièvement les difficultés rencontrées.