MU5IN852 Bases de Données Large Echelle

TME 1: analyse de la base IMDB

septembre 2021

hubert.naacke@lip6.fr

TME sur la base de films IMDB

- Explorer des fichiers de données
 - Fichier vers table
- (re)structurer les données
 - Pour une analyse :
 - Lister les attributs analysés
 - Structure du résultat des analyses
- Définir les transformations de manière déclarative
 - en SQL



Schéma de la base imdb

- Les personnes : Name, Person_Info,
- Les films : Title, Movie_Info
- Les sociétés de production : Company_name
- Les associations :
 - film/personne : Cast_Info
 - film/société : Movie_companies

Schéma d'une personne

- Une personne : Name, Person_Info
- Casting : une personne a un ou plusieurs rôles dans un film

```
Person Info (
                               Name (
                                                                         Cast_Info (
   id int,
                                🕨 id int, 🕒 🕇
                                                                            id int,
   person_id int,
                                  name string,
                                                                            person_id int,
   info_type_id int,
                                  imdb_index string,
                                                                            movie id int,
   info string,
                                  imdb id int,
                                                                            person role id int,
   note string
                                  gender string,
                                                                            note string,
                                  name_pcode_cf string,
                                                                            nr order int,
                                  name_pcode_nf string,
                                                                            role_id int)
Info_type (
                                  surname pcode string)
  id int,
   info string)
                                                                        Role Type (
                                                                           id int,
                                                                            role string)
```

Schéma d'un film

```
Title (
                                                                            Movie info
Cast_Info (
   id int,
                                         id int, ●←
                                                                                id int,
 person_id int,
                                         title string,
                                                                                movie_id int,
   movie_id int, •
                                         imdb_index string,
                                                                                info type id int,
   person role id int,
                                                                                info string,
                                         kind id int,
   note string,
                                                                                note string)
                                         production year INT,
   nr_order int,
   role_id int)
                                         imdb id int,
                                         phonetic code string,
                                                                              Info_type (
                                                                               id int,
                                         episode_id string,
                                                                                info string)
Role_Type (
                                         season nrint,
  id int,
                                         episode_nr int,
   role string)
                                         series years string)
                                      Kind_Type (
                                        id int,
                                        kind string)
```

Schéma d'une société de production

```
Movie_Companies (
                                         Company_name (
   id int,
                                         → id int,
 movie_id int,
                                             name string,
   company_id int,
                                             country_code string,
   company_type_id int,
                                             imdb_id int,
   note string)
                                             name_pcode_nf string,
                                             name_pcode_sf string)
 Company_Type (
    id int,
    kind string)
```

Datasets

- Dossier partagé : PUBLIC_DATASET
 https://nuage.lip6.fr/s/H3bpyRGgnCq2NR4
- Fichiers IMDB dans le dossier
 - imbd/vldb2015/csvfiles_sample001
- Contient aussi d'autres datasets à disposition