TP 3 - Construction d'un DataWarehouse

Ce TP a pour objectif de réaliser un data warehouse permettant d'analyser les locations de DVD d'une franchise de location appelée Sakila. Plus particulièrement, nous aborderons les aspects suivants :

- Compréhension de la base de données relationnelle (BD source)
- Modélisation du schéma en étoile à partir d'un besoin utilisateur
- Mise en place de l'ETL pour la transformation des données nécessaire à la création des dimensions et des faits
- Construction du data warehouse en utilisant Kettle.

Exercice 1 – De la BD relationnelle à la BD multidimensionnelle...

Nous disposons d'une base de données relationnelle qui recense l'activité des magasins de locations de DVD. Quelques informations utiles pour comprendre la base de données :

- Chaque magasin franchisé a constitué un inventaire des films pour la location qui est mis à jour par un des employés en fonction de locations et retours de DVD.
- Les films sont décrits par un ensemble de méta-données, dont la catégorie, le casting, la note ou la langue.
- Les clients doivent être enregistrés dans un magasin et pouvent louer des DVD dans n'importe quel magasin de la marque Sakila. La location est limitée par un délai et est payante.

La base de données relationnelle est disponible au lien suivant : http://downloads.mysql.com/docs/sakila-db.tar.gz et le schéma conceptuel est présenté en Figure 1.
Je souhaite analyser les locations de DVD en termes de nombre de locations, nombre de retours et durée moyenne de location en fonction des différents axes d'analyse :

- Du film, du client, de la date de location
- De la date de location et de retour
- Du magasin et de l'employé

Q 1.1 Dessiner le schéma en étoile permettant de réaliser cette analyse

Q 1.2 Affiner les dimensions en représentant les hiérarchies

Q 1.3 Importer la base de données relationnelles sous MySQL dans la BD "sakila"

Q 1.4 Créer la BD "sakila-wh" et créer les tables correspondant aux faits et aux dimensions. (On constituera une seule table par dimension - pas de normalisation).

Exercice 2 – Réalisation du data warehouse

Dans un premier temps, on considère que les dimensions et les faits sont gérés par des transformations indépendantes où chaque transformation inclue de principales étapes : 1) l'extraction et la transformation des données et 2) le chargement dans un élément (dimension ou fait) du data warehouse.

Q 2.1 Construire les transformations associés à chaque dimension permettant de réaliser la première étape (mise en forme des données pour être insérées dans des tables de faits ou dimensions).

Q 2.2 A la fin de chaque transformation, rajouter les étapes de chargement des données dans un data warehouse (dimension lookup/update) http://type-exit.org/adventures-with-open-source-bi/2010/07/a-basic-mondrian-cube-introducing-the-star-schema/.

Q 2.3 Créer la tâche qui permet d'exécuter l'ensemble des transformations de façon séquentielles : les dimensions et ensuite les faits.

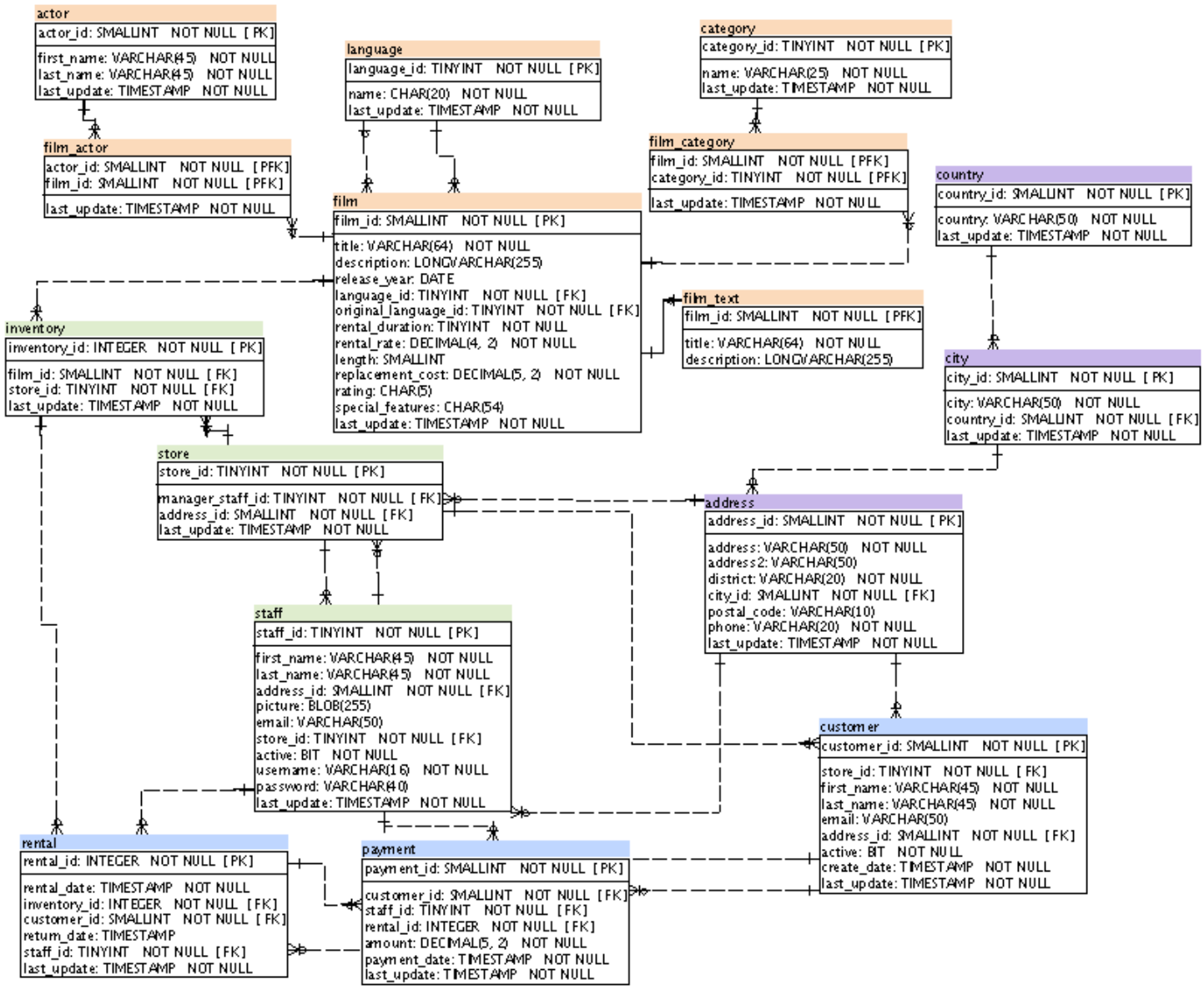

FIGURE 1 – Schéma conceptuel de la base de données Sakila

Exercice 3 – Tout ça, mais en mieux...

L'ensemble de ce process a été réalisé en prenant en compte des informations plus complexes (voir Figure 2). Les transformations et tâches sont disponibles au lien suivant : http://dac.lip6.fr/master/sakila-bi-2018/ (mot de passe donné en TME).

Q 3.1 Parcourez l'ensemble des transformations et tâches :

- Comparez avec ce que vous avez fait
- Identifier les "boîtes" non utilisées jusqu'alors et essayez de comprendre ce qu'elles font.

Q 3.2 Analyser comment sont implémentées les problématiques de mise à jour des dimensions ("Slowing changing Dimension") :

- Type 1 : "Insert/Update" (e.g., load_dim_actor)
- Type 2 : "Dimension lookup/update" (e.g., load_dim_customer)

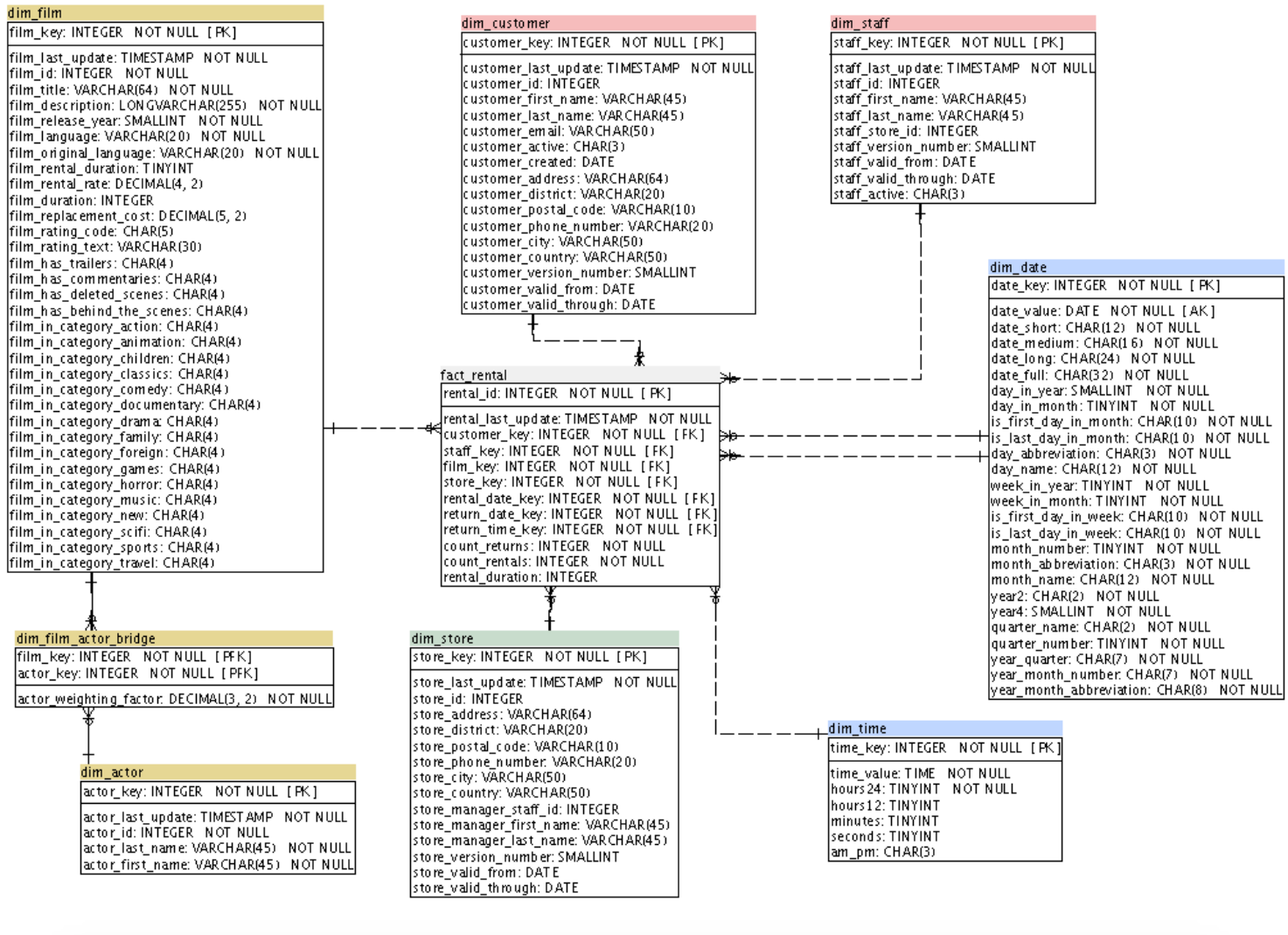

FIGURE 2 – Schéma conceptuel de la base de données multi-dimensionnelle Sakila (Modélisation complexe)