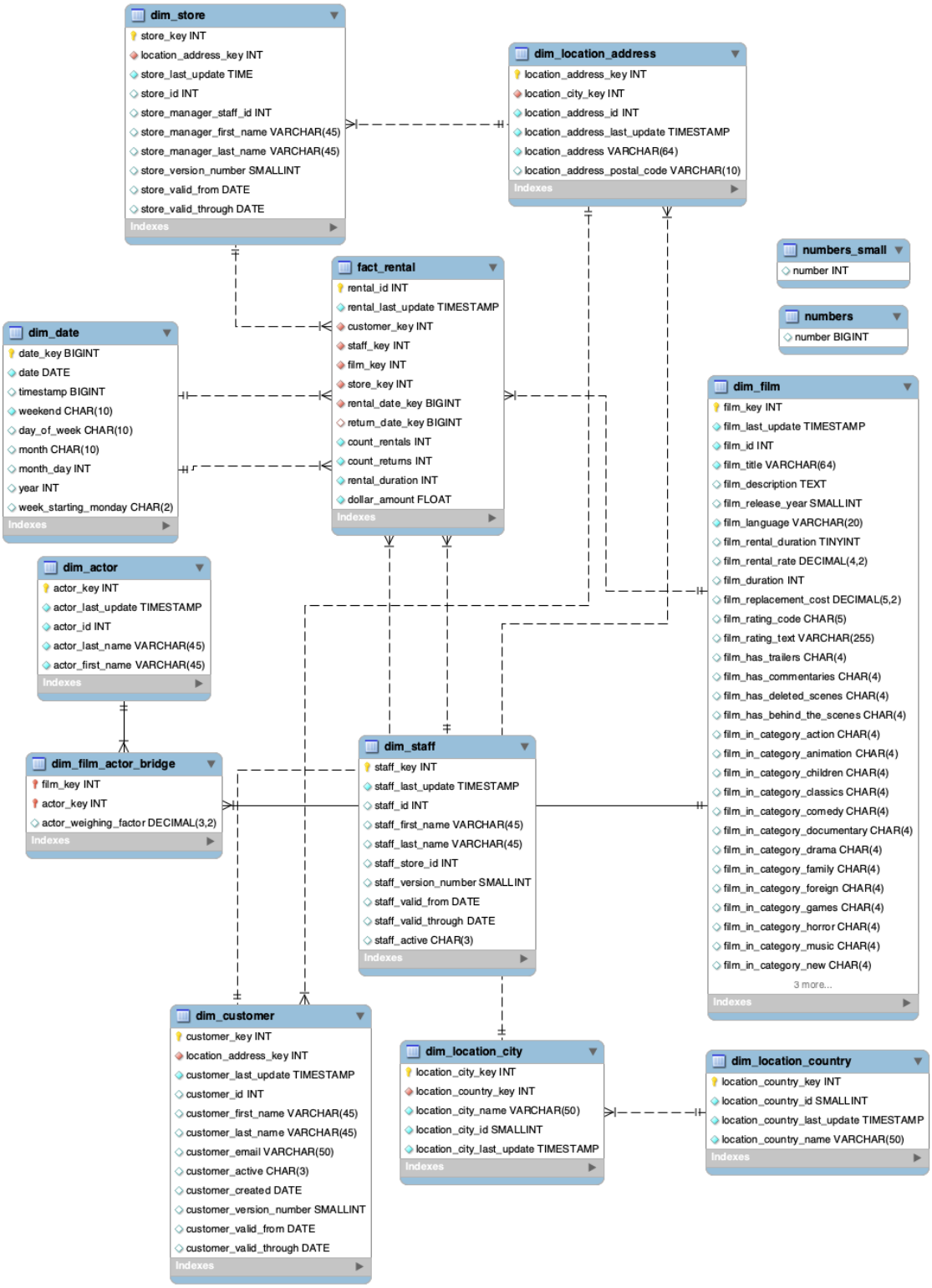
Swathi Ganesan
12372237

Assignment - 5 - ETL & Fact Table Creation

sakila_snowflake EER diagram containing fact_rental:



Swathi Ganesan
12372237

Areas for improvement in the data model:

1. *Warning: Integer display width is deprecated and will be removed in a future release.*
   The display width does not constrain the range of values that can be stored in the column. Nor does it prevent values wider than the column display width from being displayed correctly. To avoid the warnings, we can avoid explicitly specifying the display width

2. The DML insert statements into the **dim_customer** table throw warnings as we are filling values from the **create_date** (type datetime) column in the sakila schema into the **customer_created** (type date) column in the sakila_snowflake schema. We can make the datatype of the **customer_created** column to datetime to avoid warnings and potential loss of data (in this case the time information)

3. In the DML file, in multiple cases for instance, while populating the **dim_customer** table the query does not specify the schema name for the **dim_location_address** table in the FROM clause and especially in this case where we are fetching data from multiple tables and schemas it is a good practice to explicitly mention the schema names

4. From the EER diagram, we can clearly see that tables **numbers** and **numbers_small** are orphaned tables as they hold no relations to the other tables in the model. However, from the code we can see that these tables are being used for populating the **dim_date** table. We can use any increment function as an alternate and safely drop these two tables as they are useless and take up memory space

5. The **dim_film** table in the sakila_snowflake could have just a category column and that would reduce the width of the table. It makes sense to have multiple columns for the special features as we need to get rid of multi valued columns, but it doesn't make sense to do the same for category

6. The **count_rentals** and **count_returns** columns in the fact_rental table are not necessary as by definition we know that all rental ids are unique and the count_rentals column would always have the 1 for rentals

7. In terms of naming convention, all the tables and columns are named intuitively in the sakila_snowflake schema. However, the indexes created on the **fact_rental** table are named **dim_store_fact_rental_fk,**

Swathi Ganesan
12372237

dim_staff_fact_rental_fk, dim_film_fact_rental_fk, and dim_customer_fact_rental_fk respectively with fk as suffixes. I think it would make more sense to add the suffix _idx instead as this could be misleading and confusing