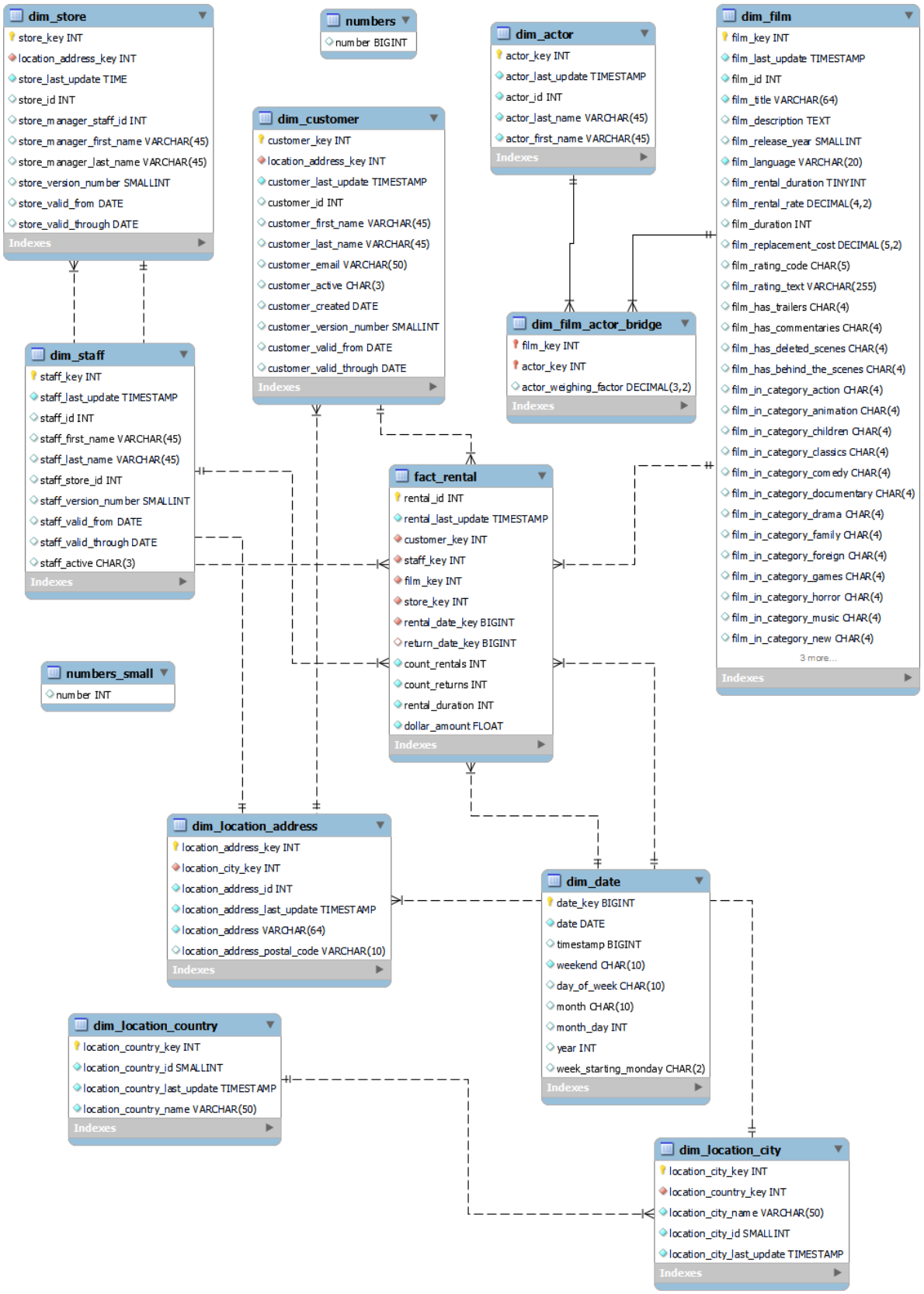
SHWETA SAMPATH KUMAR – ASSIGNMENT 5

Sakila_Snowflake EER diagram



D: Check for areas of improvement

The DDL and DML provided for the sakila_snowflake schema isn't perfect. Summarize any script errors or warnings, and areas of improvement in the data model. At minimum, answer the following:

- **Are there any warnings that should be resolved? Describe them and recommend fixes.** Received the following warnings while running the DDL & DML script:
  - *'1681 Integer display width is deprecated and will be removed in a future release.'* – This warning occurred for CREATE statements within the DDL script. Based on MySQL 8.0.19 release notes, the display width specification was deprecated for integer data types in MySQL 8.0.17 with exceptions for TINYINT (1) and ZEROFILL attributes. To avoid this error, the display width need not be mentioned for integer data types in the CREATE statements going forward.
  - *"1292 Incorrect date value: '2006-02-14 22:04:36' for column 'customer_created'"* – this warning occurred for the insertion statement into dim_customer. This is occurring because dim_customer accepts 'YYYY-MM-DD' format while timestamp format was inserted. The warning can be avoided by mentioning the date format while inserting.
- **Are there any orphaned tables that could be safely removed from the data model? Name them and explain why they can be safely dropped.**
  - Table 'numbers' and 'number_small' are orphaned tables that are not related to any of the other dimensions and can be dropped.
- **Are there any changes to column names that would help make the data model more intuitive? Describe them and recommend changes.**
  - The column dollar_amount in fact_rental could be renamed to payment_amount as it makes it more clear what the amount is for.
- **Are there any metrics that you think should be added to sakila_snowflake.fact_rental? Describe them but do not implement.**
  - payment_date could be added to check when the payment was made for the transaction
  - A calculation of the overdue charge could be added.
- **Describe any other areas of improvement you see.**
  - The dim_film table could be reduced in size by not splitting the special features and category into a new column each. If a new category or feature were to be added, then the dimensional model and table would have to be updated accordingly which in the long run would be cumbersome especially as the data increases. Even though these are binary values and they are faster in processing, it would affect performance with an extremely large data set.
  - rental_duration is not required in fact_rental table as it is already present in dim_film and can be pulled using a join when needed.
  - The column actor_weighing_factor in the dim_film_actor_bridge is not being used and isn't providing any value, so I believe it can be removed.

- The columns customer_version_number from dim_customer, staff_version_number from dim_staff, and store_version_number from dim_store could be removed as the requirement of a version number for these entities doesn't add much to the entity.
- The last_update attributes for location – country, city and address are not required as address ideally doesn't change frequently.
- The valid through and valid from attributes for customer, staff and store should be either renamed or checked if required as they are storing NULL values and not used for any business case. This is taking up additional space and can be removed.