CS 327E Final Project: Milestone 4

Prerequisites:

1. Completed Milestone 3.
2. Continuing to work with your partner.

This is the final programming milestone for the Final Project and for the course as a whole. It requires working with a new dataset and constructing an ETL process that processes this new data. This assignment sheet describes the desired output from the ETL, but it is intentionally sparse on implementation details. You are expected to choose an approach that integrates and applies the techniques we have learned in this course, implement the solution, and evaluate the results.

Dataset:

The dataset for this milestone comes from the **Cinemalytics** database which powers the web site: https://www.cinemalytics.com/. The database is a collection of Bollywood movies and music. The subset we are working with is about Bollywood Songs, Singers, and their associations.

The dataset is available for download from this link: http://cs327e-fall2017-final-project.s3.amazonaws.com/cinemalytics-with-headers.zip. Another version without header lines resides in the same S3 bucket (cs327e-fall2017-final-project) under the folder **cinemalytics**.

Desired Output:

New Tables:

- The ETL process should create 3 new tables in the IMDB database: `Songs`, `Title_Songs`, and `Singer_Songs`. A partial definition for each table is provided below:

```
Songs (song_id, song_title, song_duration)
Title_Songs(title_id*, song_id*)
Singer_Songs(person_id*, song_id*)
```

Note: An underlined field indicates a primary key and an asterisks* indicates a foreign key.

Foreign Keys:

The ETL process should create the following foreign keys:

- `Title_Songs.title_id` should point to `Title_Basics.title_id`.

- `Title_Songs.song_id` should point to `Songs.song_id`.
- `Singer_Songs.person_id` should point to `Person_Basics.person_id`.
- `Singer_Songs.song_id` should point to `Songs.song_id`.

Modified Table:

The ETL process should modify the following table:

- `Person_Basics` should be extended to include a new `gender` field.

Data Mappings:

The ETL process should map the source data to the table data as follows:

- `Songs` table should be loaded from songs.csv.
- `Singer_Songs` table should be loaded from singer_songs.csv.
- `Person_Basics` should be populated from a subset of persons.csv and singer_songs.csv based on the following criteria:
  - All new `Person_Basics` records must have a person_id value that exists in singer_songs.csv.
  - `Person_Basics.person_id` should come from the person_id column in persons.csv. This identifier is not assigned by IMDB, but it will not overlap with the existing person_id values in this table.
  - `Person_Basics.birth_year` should be extracted from the dob column in persons.csv.
  - All existing `Person_Basics` records which came from the IMDB dataset should be untouched by the ETL.
- `Title_Songs` should be populated from a subset of title_songs.csv and title.csv based on the following criteria:
  - `song_id` values should come from title_songs.csv.
  - `title_id` values should come from the imdb_id column in titles.csv.

Record Counts:

Upon completion of the ETL process, the tables should have the following record counts:

- `Songs`: 6,005
- `Singer_Songs`: 4,897
- `Title_Songs`: 5,743
- `Person_Basics`: 8,109,331

Views and Visualizations:

Upon completion of the ETL, the following database views and visualizations should be created:

- Database views that query `Songs`, `Singer_Songs`, `Title_Songs`, and `Person_Basics`.
- QuickSight visualizations based on above-mentioned views.

ER Diagram:

Upon completion of the ETL, the ERD should be updated to reflect the current state of the database based on the following criteria:

- ERD should capture any new tables since Lab 2.
- ERD should capture new columns to existing tables since Lab 2.
- ERD should capture new relationships between the tables since Lab 2.
- ERD should **not** include temp tables, intermediate/staging tables, dimensional tables, virtual views, materialized views or indexes.

Additional Notes:

- The deadline for this milestone is **Friday, 12/01 at 11:59pm**. Submit all work, including the visualizations and ERD. Follow our normal submission procedure.

- There is no starter code available for this milestone. However, you may reuse code snippets from previous milestones or lab projects.

- The M4 Grading Rubric is available from this link: http://www.cs.utexas.edu/~scohen/projects/m4-rubric.pdf