Assignment 1

PSTAT 135/235

Name: Carly Greutert

Perm Number: 8408916

MovieLens Dataset

In this assignment, we will be working on a new dataset. To download it paste the following URL into your laptop's browser: `http://files.grouplens.org/datasets/movielens/ml-latest.zip` . Alternatively, you can also go to `https://grouplens.org/datasets/movielens/` and download `ml-latest.zip` .

This dataset has around 27 million ratings on about 58,000 movies done by over 280,000 users and last updated on 9/2018. Unzip this 288 MB file. For the purpose of this assignment we will be using only two of the files that are included:

1. `movies.csv` (2.9 MB)
2. `ratings.csv` (760 MB).

Question 1: Uploading Data to BigQuery

Upload these two files into a dataset in BigQuery and call it `movie_ratings` .

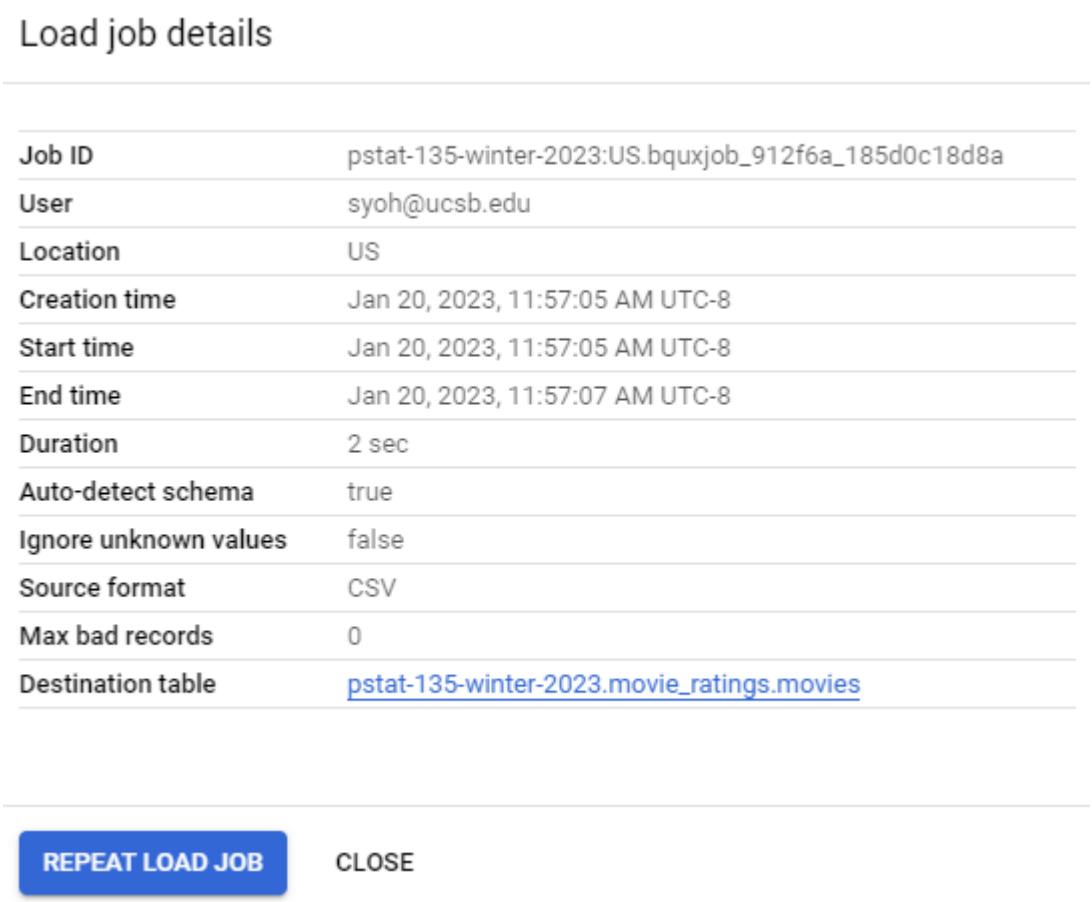
Create a new dataset and call it `movie_ratings` . We will load these two files into the newly created dataset two ways: using the web interface and agian using cloud shell.

Question 1a: `movies` table

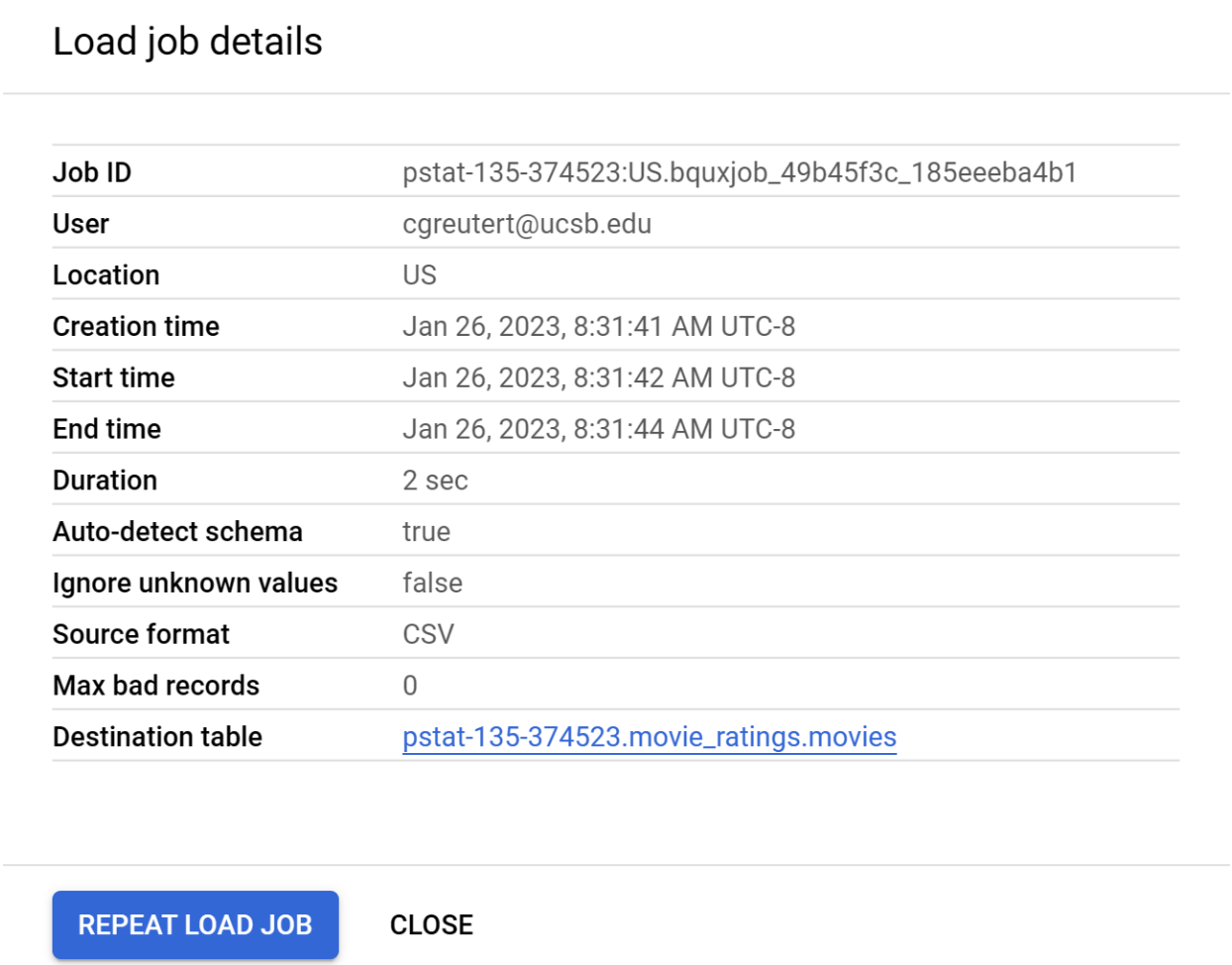
To create `movies` table from `movies.csv` file,

1. Download the zipped file
2. Unzip the archive
3. In your BigQuery interface, select in the resources list `<YOUR-PROJECT-ID>` > `movie_ratings` > click **"CREATE TABLE"** button
4. `Create table from` : `Upload`
   `Select file` : BROWSE and find `movies.csv` from your computer
   `Table` : `movies`
   Schema `Auto detect` : check

Find your LOAD job information from `PROJECT HISTORY` (next to `PERSONAL HISTORY`) at the bottom. Mine looks like @fig-job-info



Post screenshot of your LOAD job information here:



Load job details

Job ID	pstat-135-374523:US.bquxjob_49b45f3c_185eeeba4b1
User	cgreutert@ucsb.edu
Location	US
Creation time	Jan 26, 2023, 8:31:41 AM UTC-8
Start time	Jan 26, 2023, 8:31:42 AM UTC-8
End time	Jan 26, 2023, 8:31:44 AM UTC-8
Duration	2 sec
Auto-detect schema	true
Ignore unknown values	false
Source format	CSV
Max bad records	0
Destination table	pstat-135-374523.movie_ratings.movies

Question 1b: `ratings` table

Follow the same procedure as Question 1a to crate `ratings` table from `ratings.csv`. What happens?

When I try to follow the same procedure I get an error that states: "Local uploads are limited to 100 MB. Please use Google Cloud Storage for larger files." Because ratings exceeds the storage limits, we have to use a different method to upload the data from the ratings file.

PSTAT 135 Students: Upload `ratings.csv` file to Cloud Storage and create `ratings` table from it using the web interface. Then, post the screenshot of your LOAD job information here:

Load job details

Job ID	pstat-135-374523:US.bquxjob_3975a1d3_185ef28dcba
User	cgreutert@ucsb.edu
Location	US
Creation time	Jan 26, 2023, 9:38:31 AM UTC-8
Start time	Jan 26, 2023, 9:38:31 AM UTC-8
End time	Jan 26, 2023, 9:39:02 AM UTC-8
Duration	30 sec
Auto-detect schema	true
Ignore unknown values	false
Source format	CSV
Max bad records	0
Destination table	pstat-135-374523.movie_ratings.ratings

REPEAT LOAD JOB CLOSE

Question 2: `ratings` table number of rows

How many rows are there in `ratings` table?

A. 27753445
B. 27000001
-> C. 27753444
D. 27000000

There are C. 27753444 rows in the `ratings` table.

Question 3: `movies` table number of rows

How many rows are there in the `movies` table?

A. 57999
B. 58000
C. 58097
-> D. 58098

There are D. 58098 rows in the `movies` table.

Question 3: number of unique movies

How many unique `movieId` 's are in `ratings` table?

A. 52019
B. Around 27 million
-> C. 53889
D. 58097

There are C. 53889 unique movieId's in the ratings table.

What is your SQL code to obtain the info?

SELECT COUNT(DISTINCT movieId) FROM `pstat-135-374523.movie_ratings.ratings`

Question 4: highly rated movies

Which one of these movies are among top 10 highly rated movies, with at least 10,000 reviews? (select all that apply)

A. Star Wars: Episode IV - A New Hope (1977)
B. Chinatown (1974)
-> C. Godfather
D. Casablanca (1942)

What is your SQL code to obtain the info?

SELECT title FROM `pstat-135-374523.movie_ratings.movies` \ WHERE movieId in \ (SELECT movieId FROM `pstat-135-374523.movie_ratings.ratings` \ GROUP BY movieId\ HAVING COUNT(movieId) > 9999\ ORDER BY AVG(rating) DESC\ LIMIT 10)

Question 5: most watched movies

Which movie is the most watched? Make an assumption that number of ratings is strongly correlated with number of people watching it.

-> A. Shawshank Redemption
B. Forrest Gump (1994)
C. Matrix
D. Toy Story (1995)

A. Shawshank Redemption is the most watched.

What is your SQL code to obtain the info?

SELECT title FROM `pstat-135-374523.movie_ratings.movies` \ WHERE movieId in \ (SELECT movieId FROM `pstat-135-374523.movie_ratings.ratings` \ GROUP BY movieId\ ORDER BY COUNT(rating) DESC\ LIMIT 1)