Databases & SQL for Analysts

Task 3.4: Database Querying in SQL

Ngawang Dhundup

LINK NAME:

Directions:

As you've done for previous tasks, create a new text document for your answers and call it "Answers 3.4."

1. Refining Your Query: You need to get some data from the "film" table and decide to use the query SELECT * FROM film.
    - You realize that only the "film_id" and "title" columns are needed. Write a new query that selects only those 2 columns.
    - Compare the cost of the original query and the revised query, and write a few sentences explaining the comparison. Can you suggest any ways to optimize this query?

a) SELECT *

i) Seq Scan on film (cost=0.00..64.00 rows=1000 width=388)

b) SELECT film_id, title

i) Seq Scan on film (cost=0.00..64.00 rows=1000 width=19)

c) The two queries have the same cost. I believe the cost is the same because the system still has to find all of the film_id values and thus making the process similar costs. Maybe adding a LIMIT constraint, could reduce the cost.

2. Ordering the Data:
    - In the pgAdmin Query Tool, run a query that selects every film from the "film" table, with the movies sorted by title from A to Z,

- SELECT *
- FROM film
- ORDER BY title ASC

    - Then by most recent release year,

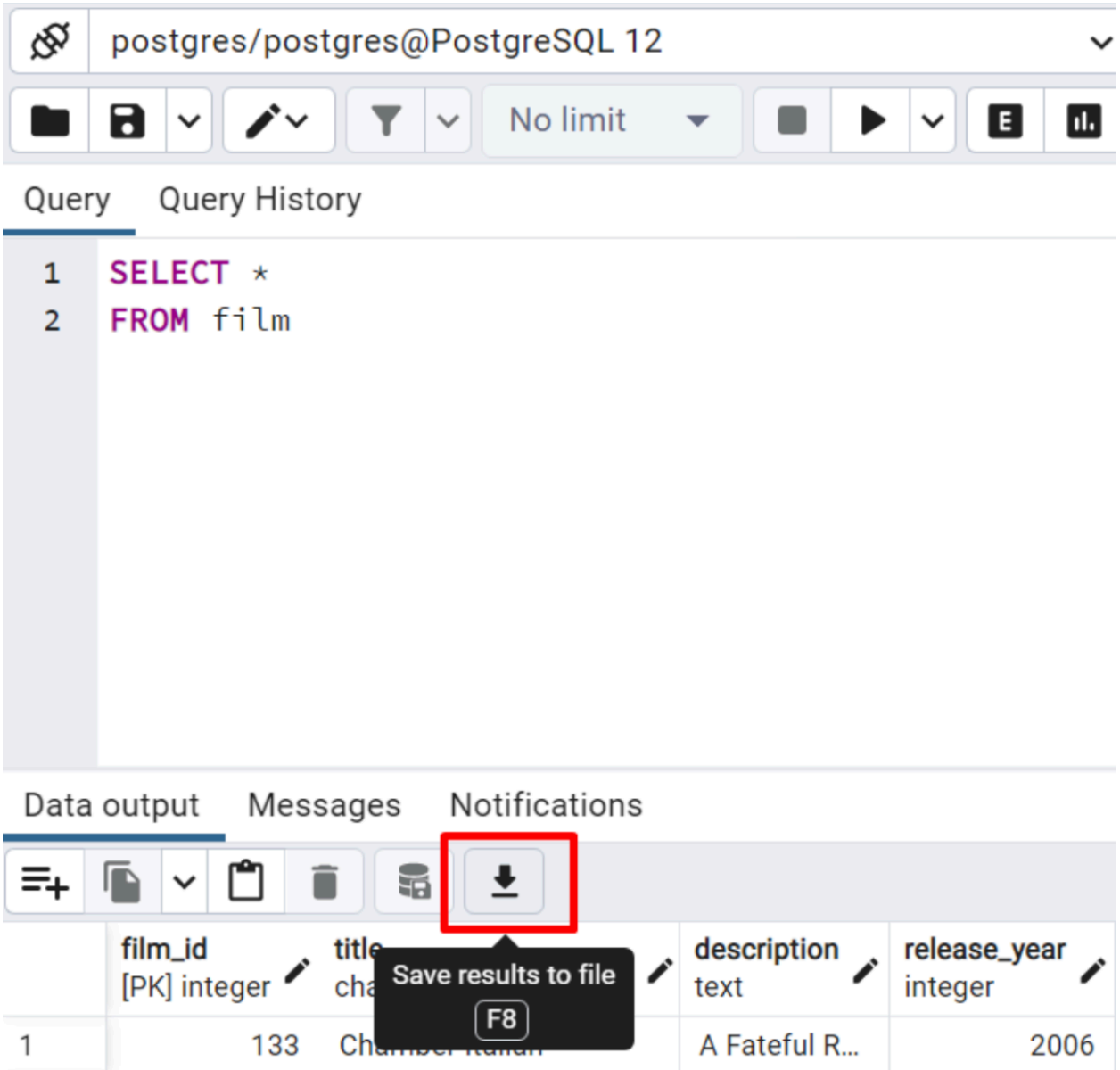
- SELECT *
- FROM film
- ORDER BY release_year ASC

    - and then by highest to lowest rental rate.

-

- SELECT *
- FROM film
- ORDER BY rental_rate DESC

    - Extract the data output of your query into a CSV file for the film collection department to analyze in Excel. To do this, click the button “Save results to file”:



3. Grouping Data: The strategy department has asked you the questions below. Write a SQL query to retrieve the correct answers, then extract your results as a CSV file.
    - What is the average rental rate for each rating category?
    -

- SELECT rating, AVG(rental_rate)
- FROM film
- GROUP BY rating
    - What are the minimum and maximum rental durations for each rating category?
- SELECT rating, MIN(rental_duration)
- FROM film
- GROUP BY rating
    - maximum rental durations for each rating category
- SELECT rating, MAX(rental_duration)
- FROM film
- GROUP BY rating

4. Database Migration: Your team has decided to use an external tool to collect data on user behavior in the new Rockbuster Android app. Data collected from this new source will need to be loaded into the data warehouse before you can analyze it.
    - Can you outline the procedure for migrating the data and who will be responsible for it?

1. **Extract** the information from the android app, typically done by the app development team or data engineers
2. **Transform** the data into the correct format (measures, dimensions, data types) so it can merge with the database correctly. Typically data engineers would also do the conversion, but data analysts should know how to as well
    - Convert dates, timestamps, and regions into standardized formats
    - Find all keys to make sure tables are linked properly
3. **Load** the new data into correct tables or create new tables and keys when necessary
    - What problems do you foresee if you start analyzing the data before it's been loaded into the data warehouse?

- If you start analyzing the data without completing the ETL process the wrong analysis could be drawn. Incorrect formats and misunderstandings could arise resulting in wrong inferences. Time will be wasted conducting multiple redundant queries, and the cost will be much higher in running complex queries.

5. Save your "Answers 3.4" document as a pdf (with screenshots) and your CSV files as a single .xlsx Excel file and upload it here for your tutor to review.