

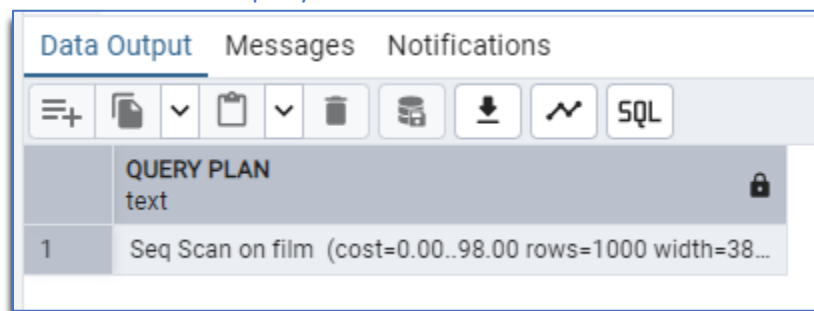
Data Immersion Ach 03.04

Ryan Wick – 01/16/2024

- **Directions**

- As you've done for previous tasks, create a new text document for your answers and call it "Answers 3.4." Make sure to include screenshots of your answers as you work through each step.

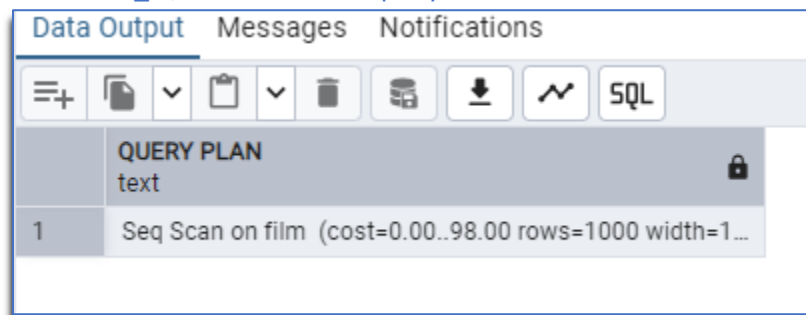
- **Refining Your Query:** You need to get some data from the "film" table and decide to use the query `SELECT * FROM film`.
 - You realize that only the "film_id" and "title" columns are needed. Write a new query that selects only those 2 columns.
 - `SELECT * FROM film` query:



The screenshot shows a database interface with tabs for 'Data Output', 'Messages', and 'Notifications'. Below the tabs is a toolbar with icons for expand, save, dropdown, copy, trash, refresh, download, and a graph. The 'Data Output' tab is active, displaying a 'QUERY PLAN' section with a lock icon. Below this, a table shows the query plan details:

	QUERY PLAN
1	Seq Scan on film (cost=0.00..98.00 rows=1000 width=38...)

-
- `SELECT film_id, title FROM film` query



The screenshot shows the same database interface as the previous one, but with the query plan for the refined query. The 'Data Output' tab is active, displaying the 'QUERY PLAN' section. Below this, a table shows the query plan details:

	QUERY PLAN
1	Seq Scan on film (cost=0.00..98.00 rows=1000 width=1...)

-
- 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?
 - From what I can tell as far as the actual "cost" of the two queries is the exact same or so small it doesn't even get reported. My guess is that they are both so similar because they not only pull a small amount of data but they should both be doing a scan in the same order. The main thing I would likely point to as an actual difference in cost is the fact that the second query only pulls two columns meaning it is far less data having to be pulled.

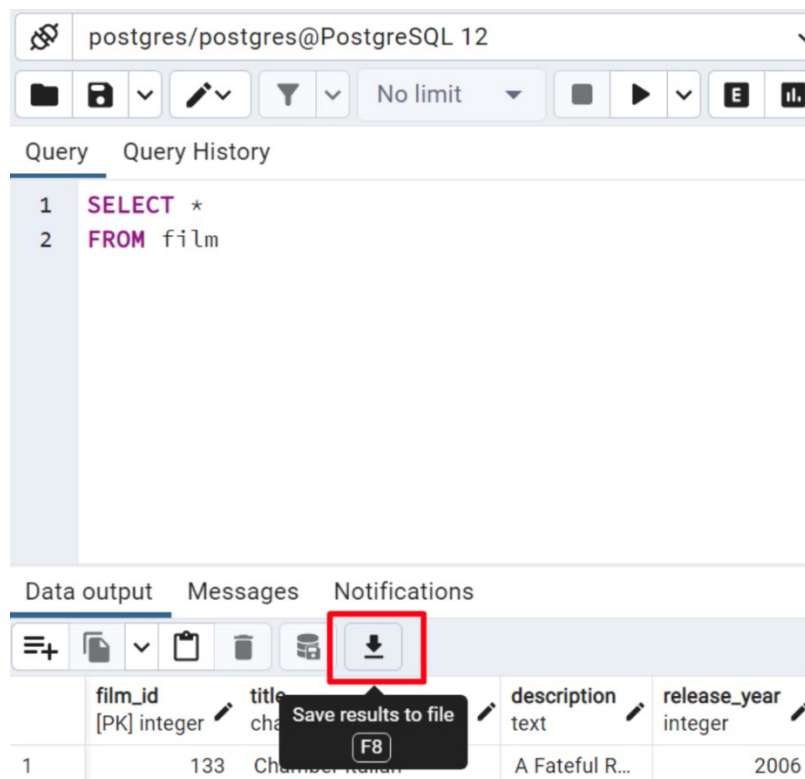
While negligible for this I can easily see how it could mean big bucks and/or time on a larger scale.

- As for optimization suggestions I would suggest using an Index and/or Where in the script to help the system more efficiently get to just the data that is desired.

- **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, then by most recent release year, and then by highest to lowest rental rate.
- 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”:

- [See attached Excel file](#)



- **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?
 - [See attached Excel file](#)
 - What are the minimum and maximum rental durations for each rating category?
 - [See attached Excel file](#)

- **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?
 - You will want to perform the Extract, Transform, Load (ETL) method. This will allow you to not only gather all the necessary information but also clean it up so there is as much uniformity as possible before loading it into a new database.
 - Typically, this would be a data engineer's tasking however depending on organizational size and/or my level of experience it may be something a data analyst would be somewhat involved in.
 - What problems do you foresee if you start analyzing the data before it's been loaded into the data warehouse?
 - Just like in our previous data cleaning exercises should we not clean the data before loading not only could it cause issues getting the database to understand the information, but it would also cause problems during the analysis. For example, if you have three different formats just for someone's birthdate alone you aren't going to get correct figures once you attempt any calculations on the data.
- 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.