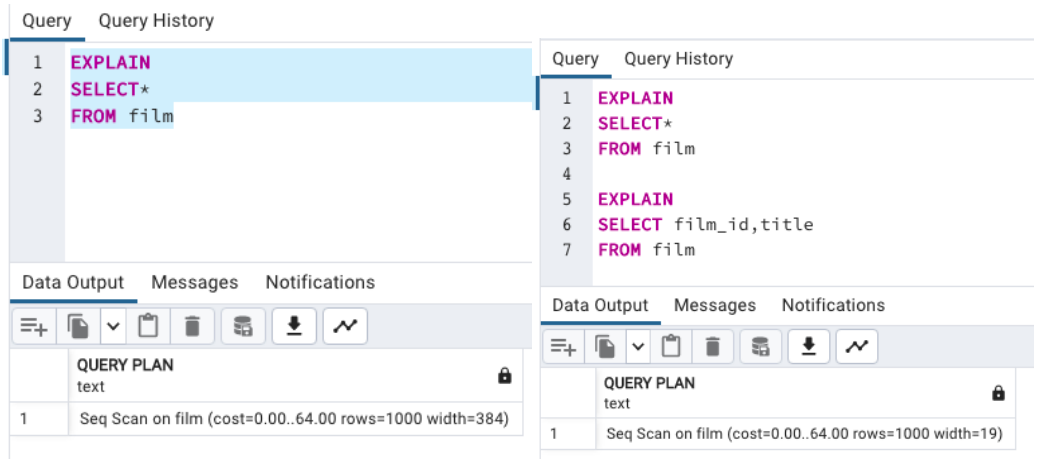
3.4: DATABASE QUERYING IN SQL

STEP 1

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?



The cost of both queries is the same, however the width is much smaller with the updated query.

The updated query helps us narrow down the columns that we are interested in and therefore show less chaotic table.

A way to optimize the query further would be to either to order the list in ascending or descending order. Or, put a LIMIT on the query to search for the TOP or BOTTOM values that interest us.

STEP 2

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.

```
SELECT title,release_year,rental_rate
FROM film
ORDER BY title ASC,
release_year DESC,
rental_rate DESC;
```

Data Output | Messages | Notifications

	title character varying (255)	release_year integer	rental_rate numeric (4,2)
1	Academy Dinosaur	2006	0.99
2	Ace Goldfinger	2006	4.99
3	Adaptation Holes	2006	2.99
4	Affair Prejudice	2006	2.99
5	African Egg	2006	2.99
6	Agent Truman	2006	2.99
7	Airplane Sierra	2006	4.99
8	Airport Pollock	2006	4.99
9	Alabama Devil	2006	2.99
10	Aladdin Calendar	2006	4.99
11	Alamo Videotape	2006	0.99
12	Alaska Phantom	2006	0.99
13	Ali Forever	2006	4.99

Total rows: 1000 of 1000 Query complete 00:00:00.579

STEP 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?
- What are the minimum and maximum rental durations for each rating category?

Query | Query History

```
15  SELECT rating,
16  AVG(rental_rate) AS avg_rental_rate,
17  MIN(rental_duration) AS min_rental_duration,
18  MAX(rental_duration) AS max_rental_duration
19  FROM film
20  GROUP BY rating
```

Data Output | Messages | Notifications

	rating mpaa_rating	avg_rental_rate numeric	min_rental_duration smallint	max_rental_duration smallint
1	PG	3.0518556701030928	3	7
2	R	2.9387179487179487	3	7
3	NC-17	2.970952380952381	3	7
4	PG-13	3.034843049327354	3	7
5	G	2.888876404494382	3	7

STEP 4

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?

  A procedure called ETL should be followed. Data engineers should be responsible for this task with support from data analysts.

  - EXTRACT - Gather the data from new Rockbuster Android APP.
  - TRANSFORM - Convert the data into suitable format.
  - LOAD - Transformed data is inserted or loaded into the new database.

- What problems do you foresee if you start analyzing the data before it's been loaded into the data warehouse?

  If we do not follow the ETL procedure, we will have to clean and analyze the data sources separately and have them in different format which will make our work less efficient.

  Having all the data collected, transformed and loaded into one singular data warehouse makes the data more accessible and allows us to work faster and more efficient.