Exercise 3.4

May 26, 2022

STEP 1

EXPLAIN

SELECT *

FROM film

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

EXPLAIN

SELECT film_id, title

FROM film

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

Looking at the two EXPLANATIONS, we see that the first part is the same, the difference lies in the "width", * = 388, while film_id & title = 19. To optimize, it's important to know what data you're looking for, this will save time.

STEP 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, 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. (You may need to explore how to save your output as a csv file in the Query Tool.)

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?

```
SELECT rating, AVG(rental_rate)

FROM film

GROUP BY rating
```

```
SELECT rating, MIN(rental_duration), MAX(rental_duration)

FROM film

GROUP BY rating
```

STEP 4

The procedure for migrating the data to the data warehouse would begin with collecting the data; this step was performed by the Android app. Once the data has been collected, the data will need to be transformed so that the data from the different sources are in the same format, for example ensuring dates are all either month/day/year or day/month/year. The final step is loading the data into the warehouse. These steps are usually completed by a data engineer.

If one were to analyze the data before it's been properly loaded, you will be working with incomplete data, this means that some things will be missing or you will receive error messages, some data may also be duplicated and skew any analysis.