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.

```
1  EXPLAIN
2  SELECT film_id, title
3  FROM film
```

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?

Data Output | Explain | Messages | Notifications

	QUERY PLAN text
1	Seq Scan on film (cost=0.00..64.00 rows=1000 width=19)

Data Output | Explain | Messages | Notifications

	QUERY PLAN text
1	Seq Scan on film (cost=0.00..64.00 rows=1000 width=388)

The difference in costs between the two is that there is a lot more space that needs to be searched when you put "*" vs "film_id, title". Being more specific lets the system know exactly what you're interested in vs returning the table to you. It's also helpful because it cuts down on the user's time spent sorting through the data to find what they want. Optimization could be done if there was something more specific you are looking for.

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.

Task 4.4 Question 2.csv

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.)

Query Editor | Query History

```
1  SELECT title, release_year, rental_rate
2  FROM film
3  ORDER BY title, release_year DESC, rental_rate DESC
```

Data Output | Explain | 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

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.**

Task 4.4 Question 3.csv

What is the average rental rate for each rating category?

Query Editor | Query History

```
1  SELECT rating,  AVG(rental_rate)
2  FROM film
3  GROUP BY rating
4  ORDER BY rating
```

Data Output | Explain | Messages | Notifications

	rating mpaa_rating	avg numeric
1	G	145251396647
2	PG	556701030928
3	PG-13	843049327354
4	R	179487179487
5	NC-17	952380952381

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

Query Editor Query History

```
1  SELECT rating,  MAX(rental_duration), MIN(rental_duration)
2  FROM film
3  GROUP BY rating
```

Data Output Explain Messages Notifications

	rating mpaa_rating	max smallint	min smallint
1	R	7	3
2	NC-17	7	3
3	G	7	3
4	PG	7	3
5	PG-13	7	3

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?

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

The data engineer would take the data from the app and then clean it up. It would be formatted in such a way that it plays nice with all the other tables already in the data warehouse. Once its clean, it'll be uploaded into the new database and able to be queried by the analysts.