

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.

Query		Query History	
1	SELECT	film_id,	
2		title	
3	FROM	film	

Data Output		Messages	Notifications
	film_id [PK] integer	title character varying (255)	
1	133	Chamber Italian	
2	384	Grosse Wonderful	
3	8	Airport Pollock	
4	98	Bright Encounters	
5	1	Academy Dinosaur	
6	2	Ace Goldfinger	
7	3	Adaptation Holes	
8	4	Affair Prejudice	
9	5	African Egg	
10	6	Agent Truman	
11	7	Airplane Sierra	
12	9	Alabama Devil	
13	10	Aladdin Calendar	

- 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?

Query

Query History

1

EXPLAIN

2

SELECT film_id,

3

title









4

FROM film

Data Output


Messages

Notifications



QUERY PLAN

text



1

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

Query

Query History

1

EXPLAIN

2

SELECT*









3

FROM film

Data Output


Messages

Notifications



QUERY PLAN

text



1

Seq Scan on film (cost=0.00..64.00 rows=1000 width=384)

Both have the same cost, as we can see but their runtimes are different. If we create a script we could save costs.

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. To do this, click the button “Save results to file”:

Query	Query History		Scra
1 2 3 4 5 6 7	<pre>SELECT title, release_year, rental_rate FROM film ORDER BY title ASC, release_year DESC, rental_rate DESC</pre>		
Data Output	Messages	Notifications	
<div><div><div><div></div><div></div><div></div></div><div><div></div><div></div><div></div></div><div><div></div><div></div><div></div></div><div><div></div><div></div><div></div></div><div><div></div><div></div><div></div></div><div><div></div><div></div><div></div></div><div><div></div><div></div><div></div></div><div><div></div><div></div><div></div></div></div></div>			
	<div>title character varying (255)</div>	<div>release_year integer</div>	<div>rental_rate numeric (4,2)</div>
1	Academy Dinosaur	2006	0.99
2	Ace Goldfinger	2006	4.99
3	Adaptation Holes	2006	2.99
4	Affair Prejudice	2006	2.99
Total rows: 1000 of 1000		Query complete 00:00:00.053	

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

Query	Query History	
1	SELECT rating, AVG (rental_rate)	
2	AS average_rental_rate	
3	FROM film	
4	GROUP BY rating	

Data Output	Messages	Notifications															
<div> <div>≡</div> <div>📄</div> <div>▼</div> <div>📋</div> <div>🗑️</div> <div>🗄️</div> <div>⬇️</div> <div>📈</div> </div> <table> <tr> <th></th><th>rating mpaa_rating</th><th>average_rental_rate numeric</th></tr> <tr> <td>1</td><td>PG</td><td>3.0518556701030928</td></tr> <tr> <td>2</td><td>R</td><td>2.9387179487179487</td></tr> <tr> <td>3</td><td>NC-17</td><td>2.970952380952381</td></tr> <tr> <td>4</td><td>PG-13</td><td>3.034843049327354</td></tr> </table>		rating mpaa_rating	average_rental_rate numeric	1	PG	3.0518556701030928	2	R	2.9387179487179487	3	NC-17	2.970952380952381	4	PG-13	3.034843049327354		
	rating mpaa_rating	average_rental_rate numeric															
1	PG	3.0518556701030928															
2	R	2.9387179487179487															
3	NC-17	2.970952380952381															
4	PG-13	3.034843049327354															
Total rows: 5 of 5	Query complete 00:00:00.070																

category?

Query

Query History

Scratch

1

SELECT rating,

2

MAX (rental_rate)

3

AS maximum_rental_rate,

4

MIN (rental_rate)

5

AS minimum_rental_rate

6

FROM film

7

GROUP BY rating

Data Output

Messages

Notifications

≡

📄

▼

📋

🗑️

🗄️

⬇️

📈

	rating mpaa_rating	maximum_rental_rate numeric	minimum_rental_rate numeric
1	PG	4.99	0.99
2	R	4.99	0.99
3	NC-17	4.99	0.99
4	PG-13	4.99	0.99

Total rows: 5 of 5

Query complete 00:00:00.038

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.

4.1 Can you outline the procedure for migrating the data and who will be responsible for it?

The procedure is called **Extract, Transform, and Load (ETL)**.

“The ETL process is:

- **Extract:** The first step involves collecting the data from multiple data sources.
- **Transform:** During this step, the extracted data is converted into another format. This could mean calculating ages from dates of birth or combining multiple data points like area codes and telephone numbers to get a contact number, for example.
- **Load:** At this point, the transformed data is inserted or loaded into the new database. “

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

There might be a problem with incomplete data format, including irrelevant and unclear data. That will result in an inaccurate analysis, time, and cost consumption for the analysis.