

Task 3.4 – Database Querying in SQL

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.
 - 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?

```
SELECT film_id, title  
FROM film
```

The screenshot shows a database query interface with a 'Query' tab selected. The query text is as follows:

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

Below the query text, there are tabs for 'Data Output', 'Messages', and 'Notifications'. The 'Data Output' tab is active, displaying a 'QUERY PLAN' section. The plan shows a single step:

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

The screenshot shows the same database query interface, but with a different query:

```
1 EXPLAIN  
2 SELECT *  
3 FROM film
```

The 'Data Output' tab is active, displaying a 'QUERY PLAN' section. The plan shows a single step:





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

The ‘cost’ of the two queries is the same, but the widths are different. The best way to optimize the query is by using scripts to run multiple queries at once.

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

Data Output		Messages	Notifications
			
	title	release_year	rental_rate
	character varying (255)	integer	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

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

Data Output		Messages	Notifications
			
	rating	average_rental_rate	
	mpaa_rating	numeric	
1	PG	3.0518556701030928	
2	R	2.9387179487179487	
3	NC-17	2.970952380952381	
4	PG-13	3.034843049327354	
5	G	2.888876404494382	

Query		Query History	
1	SELECT	rating, MIN(rental_duration) AS min_rental_duration, MAX(rental_duration) AS max_rental_duration	
2	FROM	film	
3	GROUP BY	rating	
4			

Data Output		Messages	Notifications
<div> <div>+</div> <div>📄</div> <div>▼</div> <div>📄</div> <div>🗑️</div> <div>🔄</div> <div>📥</div> <div>📈</div> </div>			
	rating mpaa_rating	min_rental_duration smallint	max_rental_duration smallint
1	PG	3	7
2	R	3	7
3	NC-17	3	7
4	PG-13	3	7
5	G	3	7

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?

Typically, data migration is done by data engineers via a procedure called ETL, or Extract, Transform, and Load.

- Extract – The data is collected from multiple data sources.
- Transform – The extracted data is converted into another format to make things uniform. For example, calculating ages from dates, or combining data points to create contact numbers.
- Load – The transformed data is loaded into a new database.

The biggest problem that I can see with data being analyzed before being uploaded into a data warehouse is unclear, mismatching data and data formats. This could lead to inaccurate analyses, as well as increasing the amount of time and cost needed to do such analyses.