Data Immersion	Achievement III		
	SQL for Data Analysts		
	Task 3.3 By Ola Gaffarova	Table of content Cover	Made for Rockbuster Stealth

DATABASE QUERYING IN SQL

Rockbuster Stealth

STEP 1A.

Query | Query History

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

Data Output | Messages | Notifications

Showing rows: 1 to 1000 | Page No: 1

	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
14	11	Alamo Videotape
15	12	Alaska Phantom
16	213	Date Speed
17	13	Ali Forever

Total rows: 1000 | Query complete 00:00:00.099

STEP 1B.

*The two queries have the same costs and rows. They differ only from their widths. The original query retrieves all columns from the film table while the revised query only retrieves specific columns. (film_id, title). Better not to use SELECT *, because it retrieves all the columns. To optimize we can use LIMIT to limit the output.*

Query | Query History

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

Data Output | Messages | Notifications

Showing rows: 1 to 1 | Page No: 1

	QUERY PLAN text
1	Seq Scan on film (cost=0.00..98.00 rows=1000 width=38...

Query | Query History

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

Data Output | Messages | Notifications

Showing rows: 1 to 1 | Page No: 1

	QUERY PLAN text
1	Seq Scan on film (cost=0.00..98.00 rows=1000 width=19)

STEP 2.

Query Query History

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

Data Output Messages Notifications

Showing rows: 1 to 1000 Page No: 1

	title character varying (255)	release_year integer	rental_rate numeric (4,2)
1	Ace Goldfinger	2006	4.99
2	Airplane Sierra	2006	4.99
3	Airport Pollock	2006	4.99
4	Aladdin Calendar	2006	4.99
5	Ali Forever	2006	4.99
6	Amelie Hellfighters	2006	4.99
7	American Circus	2006	4.99
8	Anthem Luke	2006	4.99
9	Apache Divine	2006	4.99
10	Apocalypse Flamingos	2006	4.99
11	Attacks Hate	2006	4.99
12	Attraction Newton	2006	4.99
13	Autumn Crow	2006	4.99
14	Baby Hall	2006	4.99
15	Backlash Undefeated	2006	4.99
16	Beast Hunchback	2006	4.99
17	Beauty Grease	2006	4.99

STEP 3A.

What is the average rental rate for each rating category?

```
SELECT rating, AVG(rental_rate)
FROM film
GROUP BY rating
```

Query | Query History

Data Output | Messages | Notifications

Showing rows: 1 to 5 | Page No: 1

	rating mpaa_rating	avg numeric
1	PG-13	3.0348430493273543
2	NC-17	2.9709523809523810
3	G	2.8888764044943820
4	PG	3.0518556701030928
5	R	2.9387179487179487

STEP 3B.

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

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

Query | Query History

Data Output | Messages | Notifications

Showing rows: 1 to 5 | Page No: 1

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

STEP 4.

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

The responsible person should be a Data Engineer.
The process is called ETL, which stands for Extract, Transform, Load.

1. *Extract - The engineers should first extract the data from the existing application.*
2. *Transform - Next, they need to clean and structure the data to ensure it aligns with Rockbuster's database format.*
3. *Load - Finally, they will load the transformed data into Rockbuster's database.*

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

- It could require a significant amount of time and effort to understand and clean the raw data (depending on the data size and complexity).
- *There is a higher risk of human error, especially if the process is done manually.*
- *The new data may not easily integrate with the existing dataset, potentially requiring the creation of new tables or the addition of new columns in existing tables. This would also require updates to the data dictionary and other documentation.*

Additional Note

While I believe that having more data can help clients make better decisions, it is crucial to fully understand the workflow and the structure of the incoming data in detail.
Sometimes, it may be possible to customize the data structure, naming conventions, or sorting requirements at the beginning of the process. This can save significant time and effort for data engineers in the future, ensuring that data analysts receive clean and well-structured data as quickly as possible.