Career Foundry

Data Analytics Immersion

A3.E4

Kendra Jackson

Step 1:

a. SELECT film_id, title FROM film
b. "Cost" determined using the EXPLAIN command as a prefix to the query
    1st query: "Seq Scan on film (cost=0.00..98.00 rows=1000 width=388)"
    2nd query: "Seq Scan on film (cost=0.00..98.00 rows=1000 width=19)"
    Results: Both total costs for the queries are the same. The amount of rows is the same however the width is different. The 2nd query has a shorter width, though the cost is the same this is likely the preferable choice as the result is processed faster (more time effective)
    Optimization: An index could be considered. However, the most time efficient and cost saving choice would likely be to limit the number of rows returned from the query. Depending on what the goal is with this query you could either LIMIT or TOP the results to show less results. You could also group the results using the SUBSTR command alphabetically (though this option would take more time).

Step 2:

a. SELECT title, release_year, rental_rate FROM film ORDER BY title ASC, release_year DESC, rental_rate DESC;

	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

b. CSV extraction Completed

Step 3:

a. SELECT rating, AVG(rental_rate) AS average_rental_rate

FROM film

GROUP BY rating

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

b. SELECT rating, MIN(rental_duration) AS minimum_rental_length,
MAX(rental_duration) AS maximum_rental_length
FROM film
GROUP BY rating

	rating 🔒 mpaa_rating	minimum_rental_length 🔒 smallint	maxmimum_rental_length 🔒 smallint
1	PG	3	7
2	R	3	7
3	NC-17	3	7
4	PG-13	3	7
5	G	3	7

Step 4:

a. The ETL process will be used. First the data will be extracted from the Rockbuster Android App. Then the data will be transformed to facilitate ease of analysis depending on what the company is looking for (age calculated from DOB, area code added to numbers to create a full contact number, zip codes changed to state, etc). Data is typically transformed into a standardized type which determines table, column structure and relationships (in this case, dependent on the current formatting, if any of the database warehouse). Finally, the data would be loaded into the current data warehouse. Typically, a data engineer is responsible for the ETL process, that being said sometimes analysts may take part.

b. Data formatting may not match, data could also be null or blank because constraints have not been applied. You may also have duplicates. Data is in essence incomplete. While you may have some analyzable data, you are not analyzing to the full scope of the data and therefore do not have an accurate analysis.