Answer 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.
   * 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?

1a.

SELECT film\_id, title

FROM film

Graphical user interface, text, application

Description automatically generated

1b.

From the data output message both the original and the revised query would be the same cost(0.00…64.00). Other than the rows and widths being different which are their outputs and runtime being different there prices would stay the same. A way to better optimize these queries would be to create a better script that the RDBMS can more easily use to manipulate the data asked for within the data warehouse.

1. **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”:

*1a.*

*Table

Description automatically generated*

*1b.*

*CSV. File attached in EXCEL Sheet Tab “ORDER”*

1. **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?

Screenshots of queries are below:

Excel file sheet Tab “AVG”

1a.

Graphical user interface, table

Description automatically generated

1b.

Max/Min

Excel file sheet Tab “MIN MAX”

Graphical user interface, text, application

Description automatically generated

1. **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?

1a.

To start off we want to migrate user behavior from the new Rockbuster Android app to the Rockbuster rental movie database through a procedure called ETL (Extract, Transform, and Load). The first step “Extract” involves collecting the data from the multiple data sources in the new Rockbuster Android app. During this step, the extracted user behavior data is converted into another format. This could mean calculating the length of user’s rental duration for rental rate. At this point the transformed data is inserted or loaded into the new database. Data engineers are responsible for these types of duties, but a data analyst should know the logistics of the ETL procedure to better communicate issues with data engineers.

1b.

Problems that may arise if data is analyzed before its been loaded into the data warehouse can be quality issues. Formatting of the data may not match the data that is housed in another database. But RDBMS are used to fix possible issues that come up with faulty data.