Answers 3.4

* **Refining Your Query**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

* The cost for both the original query (**SELECT \* FROM film**;) and the revised query (**SELECT film\_id, title FROM film**;) remains the. However, the revised query has a significantly smaller width (**19 vs. 384**), meaning it processes less data by retrieving only two columns instead of all.
* **To optimize the query**:
* Avoid using **SELECT \*** to reduce data retrieval overhead.
* Create an index on the film\_id and title columns to improve query performance, especially for large tables.
* **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.

****

* **Grouping Data**
* What is the average rental rate for each rating category?

A screenshot of a number of rating

Description automatically generated

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

A screenshot of a computer

Description automatically generated

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

1. **Extract Data**: Use the external tool to collect user behavior data and export it in a standardized format (e.g., CSV, JSON).
2. **Transform Data**: Clean and preprocess the data to align with the schema of the data warehouse. This may include deduplication, data type conversions, and validation.
3. **Load Data**: Import the transformed data into the data warehouse using ETL tools like Talend, Informatica, or native database tools.
4. **Verification**: Validate the loaded data to ensure accuracy and consistency.
5. **Responsibilities**:
   * **Data Engineers**: Oversee the ETL process and ensure the pipeline is robust.
   * **Data Analysts**: Validate data quality post-migration.
   * **IT Team**: Handle any infrastructure or tool-related issues.

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

1. **Incomplete Data**: Analysis might be based on partial data, leading to inaccurate insights.
2. **Inconsistency**: Without preprocessing, the data may contain duplicates or errors, reducing reliability.
3. **Performance Issues**: Accessing raw external data directly can strain systems and increase latency.
4. **Integration Challenges**: Data might not align with existing datasets, complicating analysis.

* **Bonus Task**

**A screenshot of a computer

Description automatically generated**