DTI 5126: Fundamentals for Applied Data Science

Summer 2023 Assignment 1

Submission Deadline: 17th June, 2023 on Brightspace.

This assignment should be completed by a team of 2 students.

The assignment is in two parts: SQL & Data Warehousing. Upon completion, present your result as a PDF document (not more than 10 pages), including the answers generated or plots. Where applicable, submit the source codes used to generate your results as a separate attachment using <LastName_FirstName>.SQL or <LastName_FirstName>.R extensions. **Do not zip your file submissions.**

Part A: SQL using PostgreSQL Server (50 points)

You are hereby provided a list of flights that occurred in the first month of 2015 along with other flight information. Using the SQL scripts and data provided, create a database named *Flight_System* and insert data into the tables to populate it. Write SQL queries to complete the following:

a. Provide the key summary statistics of the data contained in the table by retrieving the number of distinct aircrafts, total number of flights as well as a few statistics about flights departure delays (e.g., min, max & avg departure delays). **(5 Points)**
b. Create a view called *FlightSummaryView* to display the date (e.g., 2015-01-01), iata_code, origin_airport, concatenated city, state and country renamed as *Address*, and the total number of flights departing from each airport for the first week of 2015. Use the JOIN ON syntax and order by the iata_code in descending order (Make sure to add space between the address if required). **(5 points)**
c. Display the origin_airport, destination_airport, and the rank for the top 3 routes departing from each airport. **(5 Points)**
d. Display the airport iata_code, airport name, airline iata_code, airline name, flight_number, tail_number, origin_airport, destination_airport, departure_time, and arrival_time for all flights that fly on weekends (Saturdays and Sundays) and landed between 4 and 5 am. **(5 Points)**
e. All New York flights originate in one of 3 airports: 'JFK' (Kennedy), 'LGA' (La Guardia), and 'EWR' (Newark in New Jersey). Count how many flights originate at 'JFK.' Then show how many flights originate at 'JFK' as a percentage of all flights. (hint: use a WITH clause or a FROM subquery). **(10 Points)**
f. Retrieve the flight information for all flights going into New York City, flying through any of its two airports (JFK and LGA) or into neighboring city's airport New Jersey (Newark, EWR), where the elapsed time is greater than 500 mins. Suppose we are told these flights are cancelled. Use this information directly in SQL to update their cancelled status from 0 to 1. **(10 Points)**
g. Build a single temporary table called *Departure_Delays* that capture the categories of the departure_delays of flights based on how many are 'big,' 'medium,' and 'small' delays. Provide the iata_code, airline, departure delay category, and determine the total number of delays in each category. Order the result based on the total number of delays in descending order. **(10 Points)**

Part B: Data Warehousing & OLAP (50 points)

H.S. designs is an interior design company that specializes in home kitchen designs. The company offers a series of seminars for free at home shows, kitchen and appliance stores, and public locations as a way to build its customer base. The company earns revenues by selling books and videos that instruct people on kitchen designs. They also offer custom-design consulting services. The company has a database that keeps track of its customers, the seminars

they attended, the contact details, and the purchases made. H.S. Designs will like to build a data warehouse to analyze the sales of its products. The fact table for such a data warehouse might be:

Sales (TimeID, CustomerID, ProductNumber, Quantity, UnitPrice, Total)

The *TimeID* points to the Timeline dimension table with the attributes (TimeID, Date, Month_text, e.g. october, Quarter_text, e.g., Qtr 3, Year). The *customerID* points to the Customer dimension table with the attributes (CustomerID, CustomerName, Email, PhoneAreaCode, City, State and ZIP). The *ProductNumber* points to the Product dimension table with the attributes (ProductNumber, ProductType and ProductName). The *Quantity* attribute is the number of seminar ordered, the *UnitPrice* is the cost and the *Total* is what the customer paid. Using the SQL scripts provided, build a data warehouse for H.S. Designs named HSD_DW and insert data to populate the tables.

Deliverables:

1. Sketch a representative Star schema for the data warehouse (specifying the relations, the attributes, the primary keys, and the foreign keys). **(15 points)**
2. Suppose that we want to examine the data of HSD_DW, write SQL queries to answer the following questions: **(20 points)**
    a. Which customer(s) made an order in the past 90 days from May 31, 2018? Provide the CustomerName and CustomerID, Quantity and Total amounts of the orders.
    b. Which customer had an average order greater than the average order of all customers?
    c. For each customer, determine the time between the sale of products as *Days_between_Product_Sales*. Display the Customer ID, Customer Name, Product Number, Product Name, Date, End Date, *Days_between_Product_Sales*. Consider using the lag function and order the result by the CustomerID.
    d. Write SQL query for the ”Roll-Up” operation to summarise the total sales per quarter.
3. Customer churn is a huge problem for telecoms providers. Analyze the *customer_churn* dataset provided to determine ways to improve customer retention. Using Excel or R, build an OLAP cube to determine the following: **(15 points)**
    a. The *total revenue* contribution from a *Two Year* contract for each *Offer* by *internet type*.
    b. For *Offer B*, what (%) of the *total revenue* was contributed by *churned customers* that accepted a *Month-to-Month* contract for *Cable* service.