DTI 5126: Fundamentals for Applied Data Science

Summer 2022 Assignment 1

Submission Deadline: 9th June, 2022 on Brightspace.

This assignment should be completed by a team of 2 students using MS SQL Server & R.

The assignment is in two parts: Database & 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: RDBMS & SQL (40 points)

VRG is a smart art gallery that acquires and sells contemporary fine art, including original paintings, prints and other artwork. VRG also provides art framing services and sells books on art and artists. VRG has been in business for over 30 years and has one full-time owner, 3 sales reps and 2 workers that make frames, hang art in the gallery and prepare artwork for shipment. Please note that VRG is an acquisition database and has a sales order system that interfaces with the acquisition database. Using the SQL scripts and data provided, create a database named VRG in MS SQL Server and insert data into the tables to populate it. Write SQL queries to display the following:

a. Identify transactions with null values on the DateSoldID and remove them from the table **(5 points)**.

b. List the WorkId, Title, Medium, ArtistID, and the concatenated artist name renamed as FullName for all artwork that the title contains the word "Yellow, "Blue" or "White", e.g., the title "On White II" would meet the criteria **(5 points)**.

c. For each Artist, show the Year, ArtistID, sum of SalesPrice as *SumOfSubTotal*, and average of SalesPrice as *AverageOfSubtotal* for each year. **(5 Points)**

d. Show the ArtistID , FirstName, Lastname, WorkID, and Title of Artists that have an artwork sold with a SalesPrice above the average SalesPrice **(5 Points)**

e. Modify the email of the customer Johnson Lynda and her encrypted password from NULL to Johnson.lynda@somewhere.com and "*aax1xbB*" respectively. **(5 Points)**

f. For each customer, find the time (in days) between a purchase and the next for the DateSoldID. Display all the attributes of the customer and days between purchase as *Days_Difference*. Consider using the Lead or Lag function. **(5 points)**

g. Create a view called *CustomerTransactionSummaryView* to display the concatenated customer name renamed as *FullName* using the LastName and FirstName, Title, DateAcquired, DateSold, and difference in the AcquisitionPrice and SalesPrice as *Profit* for art works with an AskingPrice greater than $20,000. Use the JOIN ON syntax and order by the AskingPrice in descending order (Ensure to add space between the full name if required). **(5 points)**

h. Build a single temporary table called *Purchase* that captures customers' purchases from 2015 to 2017. The table should contain the TransactionID, DateAcquired, CustomerID, LastName, FirstName, first AcquisitionDate as *MinAcquisitionDate*, last AcquisitionDate as *MaxAcquisitionDate*, and Medium used for the artwork. Also, the Medium values should be represented as numeric values using High Quality Limited Print – 1, Color Aquatint – 2, Water Color and Ink – 3, Oil and Collage – 4, Others - 5. Note: consider using CTEs and CASE statement in your query if required. **(5 Points)**

Part B: Data Warehousing & OLAP (60 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 snowflake 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 **(15 points)**
    a. Write an SQL query to answer the following question: ”Which customer(s) made an order containing at least five products with different product numbers?” Provide the CustomerName and CustomerID
    b. Write an SQL query for the following report: ”Which customer(s) made the largest order, i.e., those that would result in the largest bill?”
    c. Write SQL queries for the ”Roll-Up” operation to summarise the total sales per Year.
3. Suppose an analyst finds that monthly total have decreased from April 2018 to June 2018, instead of growing. The analyst wishes to check if there are specific product type or customer city that are responsible for the decrease **(15 points)**
    a. What are the aggregates that the analyst would start with?
    b. What are the relevant “drill-down” operations that the analyst would need to execute?
4. Using R, read the dimensions files and the Product_Sales fact table. Build an OLAP cube for the Sum of Total Quantity. **(15 points)**