Task N: Advanced database  
No of words of report is 900-1000.

Produce a word-processed answer to this assignment.

Please use Arial font and a font size of 12 for text.

For SQL code and output, use courier new font and a minimum size of 10 to preserve the SQL format. Where necessary, screenshots of SQL output should be used.

**You must not run the sh\_idx.sql and the sh\_cremv.sql scripts on your DWU instance. They are provided for information purposes only and as a reference to help you avoid recreating the same indexes and MVs that already exist in SH2.**

**Background and Scenario**

**This assignment is based on two scenarios as follows:**

a)

The Sales History (SH) Data Warehouse scenario. SH is a sample database schema provided by Oracle, which has been extensively used in the Oracle’s Data Warehousing Guide (Potineni, 2021). The details of this scenario are provided in Appendix 1.

b)

GLOBAL CREDIT CARDS scenario of their credit card customers defaulting on their due credit card payments as described in Appendix 2.

**Assignment Questions**

**Part 1: Data Warehousing Tasks (50 Marks)**

**This part is based on the Sales History scenario as described in Appendix 1.**

**Provide all the SQL queries and any other code that you wrote in answering any of the tasks / questions (e.g., the use of EXPLAIN PLAN statements for the queries and their outputs using Spooling or other suitable means).**

(A)

-Study the index definitions in sh\_idx.sql. These indexes have already been created in SH2. Whatever indexes you decide to create for this task should be the result of your own research and thinking, and MUST be different from those already existing in SH2 or those indexes defined in the Oracle Data Warehousing Guide (Potineni, 2021) or those of other students.

-You need to design two queries such that each query involves at least three different tables and at least one aggregate function. You need to ensure that your queries have adequate selectivity such that if suitable indexes were available in your personal version of the database, the queries would have performed more efficiently.

-You need to identify and justify at least two indexes to improve the performance of your queries. Then create your proposed indexes in your DWU version of the database. You need to run your queries before and after creating your proposed indexes and report EXPLAIN PLAN outputs and make sure that your proposed indexes have been used by your queries and have **improved their performance significantly**.

-Then critically discuss the differences in the performance of your queries with and without the proposed indexes. **You need to critically review and cite relevant database literature to support your choice of indexes** and how you dealt with the issue of selectivity in your queries.

(B)

-There are two materialised views (MVs) defined in sh\_cremv.sql and these MVs have already been created under SH2 shared schema. You should study these two MVs and understand their benefits to the user of the SH2 data warehouse.

-You then need to design and create two new MVs on the base tables in your personal schema. Each of your proposed MVs should involve at least three different tables and at least one aggregate function. Justify why these two new MVs would be useful for the users of your data warehouse. Note that you must create brand new and unique MVs, based on your own research and thinking, and these MUST be completely different from those of SH2 or those MVs defined in the Oracle Data Warehousing Guide (Potineni, 2021) or those of other students.

-Then design two queries such that when you run these queries, the database optimizer will re-write these queries and instead of the tables named in your queries, the system will use the two new MVs to answer the queries. Note that the queries should return subsets of the values contained in these MVs.

**You must not query your MVs directly in the FROM clause; let the database optimizer re-write these queries and answer them using the new MVs.**

-You need to run your queries on both the SH2 schema and on your personal schema and report EXPLAIN PLAN outputs. You should make sure that the queries on the personal schema use the new MVs and have significantly better performance compared to the same queries’ performance when run on the SH2 data warehouse as the newly proposed MVs would not exist in the SH2 schema.

-Then critically discuss the differences in the performance of your queries with (in the case of your personal schema) and without (in the case of SH2 schema) the proposed MVs.

**You need to critically review and cite relevant database literature to support your choice of MVs and queries**.

(C)

-Choose, justify, apply, and critically assess application of dimension objects to your personal schema.

**Note: Your personal version must have an instances starting with the prefix DMU59. Example,** SELECT \* FROM **DWU59.**CHANNELS**;**

**Part 2: Data Mining Tasks (35 Marks)**

**This part is based on the GLOBAL CREDIT CARDS company’s credit card customers scenario as described in Appendix 2.**

**The main purpose of this part is to correctly predict if credit card customers will default on their due payments. You are required to perform the following tasks:**

A.

Explore the dataset and justify whether GLOBAL CREDIT CARDS company’s problem belongs to predictive or descriptive data mining models. Choose, WITH JUSTIFICATION, which data mining task (e.g., classification, association rules, clustering, regression, etc) will be used to produce data mining models for the GLOBAL CREDIT CARDS company’s scenario.

B.

Prepare and set up your views and tables under your personal account for accessing the shared UnitedCreditCards dataset, which also includes splitting the dataset for building, testing and applying the data mining models.

C.

Using the PL/SQL Data Mining API, develop at least TWO models using suitable algorithms for performing your chosen data mining task on the UnitedCreditCards dataset.

D.

Using suitable metrics, evaluate capabilities of the models you have developed for this task.

E.

Present and critically discuss your findings and make recommendations to the Managing Director of GLOBAL CREDIT CARDS company.

**Part 3 (15 marks)**

**Critically evaluate the SH data warehouse and the GLOBAL CREDIT CARDS company’s** UnitedCreditCards **dataset in relation to the theory and best practices of data quality and standards.**

The report should be concise and comprehensive and in the region of 900-1000 words. You should use Harvard style of citation and referencing by following the guidelines in Pears and Shields (2008).

**Appendix 1**

**The Sales History (SH) Data Warehouse**

SH is a sample database schema provided by Oracle, which has been extensively used in the Oracle’s Data Warehousing Guide (Potineni, 2021). The SH schema, as shown in Figure 1, consists of a big fact table, SALES, and five relatively small dimension tables: TIMES, PROMOTIONS, CHANNELS, PRODUCTS and CUSTOMERS. The additional COUNTRIES table linked to CUSTOMERS creates a simple snowflake. The model and the attributes are aimed at demonstrating data warehousing functionality like star transformation and query rewrite. They do not necessarily represent the optimal approach for this kind of data warehouse in real productive environments; and such a design would be driven more by business requirements than by the star itself. Table 1 shows the cardinalities of the tables in the SH schema.

| **Table** | **Number of Rows** |
| --- | --- |
| Customers | 50,000 |
| Countries | 19 |
| Products | 10,000 |
| Sales | 10,16,271 |
| Costs | 7,87,766 |
| Times | 1,461 |
| Promotions | 501 |
| Channels | 5 |

Table 1: Cardinalities of SH tables (as per SH2 implementation).

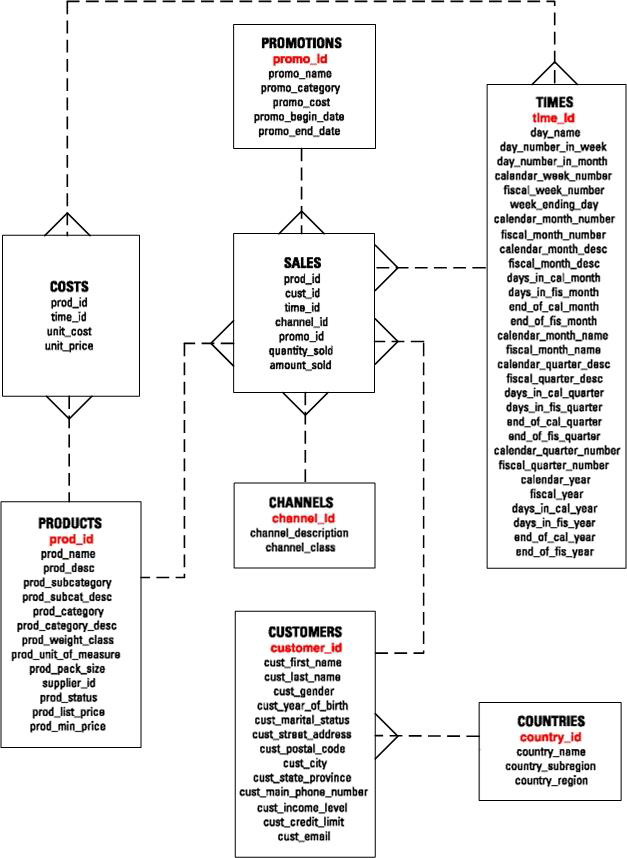
****

Figure 1: SH Star Schema.

Usually, a star schema includes a single fact table and few dimensional tables, however, the SH schema designer identified another useful fact table called COSTS, which is linked to the TIMES and PRODUCTS dimensional tables.

**Two versions of SH schema/database**

1)

The optimised version of the SH database is created under the SH2 user. You have read-only access to query any of the tables of SH database by prefixing any table or view name by “SH2.”, e.g., to query the contents of CHANNELS table, use:

SELECT \* FROM SH2.CHANNELS;

2)

An un-optimized version of the SH database has been created under your own username DWU59. Throughout the first part (data warehouse), you MUST use this special username.

**Important Note**

**Whenever, you need to assess the performance of existing database structures (e.g., materialised view, index) you must refer to SH2 as above (using the “SH2.” prefix). Any new database structures that you need to create should be created under your username DMU59. Likewise, when you need to assess the performance of any new database structures that you create under your DWU username (DWU59), you need to prefix all tables/views/materialised views accordingly, e.g., to query the contents of your CHANNELS table, use:**

SELECT \* FROM **DWU59.**CHANNELS**;**

**Appendix 2**

**GLOBAL CREDIT CARDS company’s credit card customers scenario.**

**Dataset:** GlobalCreditCards, a shared read-only table in the Oracle Database CISBG

**Response variable: Default payment indicator,** defaultnm

**Data dictionary:** Data\_Dictionary.xlsx

GLOBAL CREDIT CARDS is a hypothetical credit card company operating in Nigeria and most of the Europe and Commonwealth countries. GLOBAL CREDIT CARDS is having issues with many of its customers (1 in 5) who are defaulting on their due payments, and causing the company to lose money, customers and reputation. In an effort to reduce the number of customers defaulting, the company wishes to develop a model that will allow them to better predict if a customer will default within the next month, if so then to offer suitable terms to the customers that may allow them to avoid defaulting on their payments. The task is to develop a few models and choose the one that can predict with highest possible accuracy on whether a customer will default on their next monthly payment.

**Important Note**

**You need to prefix all instances of the data mining part with DWU20**

***Note:***

Code files need to be provided. All codes should have comments explaining the lines of code used.

The following use of AI tools or software engineering automation (SEA) tools will lead to a score of zero:

•code, data or text generated by AI & SEA tools

•diagrams or models created by AI & SEA tools either from assignment specification text or reverse engineering or mapping code to diagrams

•any data dictionary descriptions or definitions or specifications generated by AI & SEA tools