**ST1501 CA2 Group Tasks**

**Class: DAAA/FT/2A/03**

**Group No: 6**

**Group Members: Loh Yip Khai, Koh Zi Hao, Isaac Low**

|  |  |  |
| --- | --- | --- |
| **Student No** | **Name** | **Team Lead (Y/N)** |
| **P2317454** | **Loh Yip Khai** | **Y** |
| **P2317483** | **Koh Zi Hao** | **N** |
| **P2342348** | **Isaac Low Zu Le** | **N** |
|  |  |  |

**Group Tasks Solution**

1. Submit OLTP\_insert.sql including SQL queries that insert data into the OLTP tables.

Identify data quality issues in dataset provided. List down details in the table below.

|  |  |  |
| --- | --- | --- |
| Issue No. | Description of issue (which file, which line/part, what is the problem) | Solution (how do you resolve the issue) |
| 1 | modeling.xlsx sheet\_name=’great ideas’  Irrelevant to the project | Did not include this sheet |
| 2 | Modeling.xlsx sheet\_name=’modeltype’ and “model”.  “LogR” in modeltype sheet’s “Model Code” column is not consistent with “logR” in the model sheet’s “Model Code” column | "LogR" changed to "logR",  under Model Code column in Modeling.xlsx, sheet\_name”modeltype” |
| 3 | Orders.csv  1 Missing value(Blank) and 1 incorrect DataType(c0068xxxxxxxx) for CustomerID | Removed row where CustomerID is (Blank)  Removed Trailing “xxxxxxxx” for CustomerID(c0068xxxxxxxx) to become (c0068) |
| 4 | Employee.docx  There are 73 records of the contact number that is found duplicated | Kept It as it is as we cannot drop these records |
| 5 | Customer.csv  There are 4 records of the contact number that is found duplicated | Kept It as it is as we cannot drop these records |

Explain how you check if your table creation and data insertion are correct.

We first created the tables with the columns and refer to the given dataset to verify whether the column names are correct. We also used SELECT \* FROM [table] and referred to the given dataset file to check the few 10 rows to ensure that data insertion is correct.

1. Paste your data warehouse design here (database diagram). Write a short description to explain your design such as choice of measurement and levels of details.

|  |  |
| --- | --- |
| DW design | Show all details of the DB diagram. Format your layout and size of tables to ensure nothing is hidden. |
| Explain your design | The database is using a snowflake schema,, with a fact table in the center and different dimensions like employee dimension, time dimension, customer dimension and model dimension surrounding I, and a dataset and model type dimension branching out from the model dimension. We chose to use a snowflake schema in order to reduce data redundancy and as snowflake schemas offer better data consistency and uses less space. |

1. Implement the data warehouse you designed in b) using MS SQL server.

Submit DW\_create.sql with all the SQL statements that creates DW tables, and DW\_insert.sql with all the SQL statements that query data from OLTP tables and insert into DW tables.

Briefly explain how you verify that your data warehouse is setup correctly, and that the data is inserted correctly.

-We ensure that each Dimension in the Data Warehouse has the same number of rows of the OLTP database for the respective tables(Customer, Employee, Model, ModelType, Dataset).

-After creating the empty data warehouse, we check that the empty Dimensions have the correct Column names(Eg. for Customer\_Dimension,   
CustomerSK, CustomerId, FirstName, LastName, CompanyName, Contact all exists)

-After inserting Data, we check that the number of rows affected matches with the rows for the respective OLTP Tables, to ensure that all the data is loaded into the Datawarehouse.

1. Implement the queries and explain your findings to the 3 questions. You can list no more than 3 findings/queries for each question below. Modify the template accordingly.

* Q1: (Query 1)

What do you want to find out?

The goal of the analysis is to understand the trends in profits and order volumes over time. Specifically, we want to identify any seasonal patterns, periods of significant growth or decline in profits, and how the number of orders correlates with total profits across different quarters and years. This information can help in strategic decision-making, such as resource allocation, marketing efforts, and forecasting future profits.

Insert your query here:

SELECT

t.[Year],

t.[Quarter],

SUM(f.Price) AS TotalProfit,

COUNT(f.OrderSK) AS TotalOrders

FROM

SPAIDW2A0306.dbo.fact\_table f

JOIN

SPAIDW2A0306.dbo.time\_dimension t ON f.TimeSK = t.TimeSK

GROUP BY

t.[Year], t.[Quarter]

ORDER BY

t.[Year], t.[Quarter];

Insert your results here:

A screenshot of a data

Description automatically generated

Explain what you find based on the results:

**Seasonal Peak in Q4**: The fourth quarter consistently shows the highest profits and number of orders, suggesting a strong seasonal trend likely tied to holiday spending or end-of-year budgeting.

**Year-over-Year Growth**: There is a clear upward trend in both profits and order volumes from 2021 to 2023, indicating robust business growth.

**High Correlation Between Orders and Profits**: Increased order volumes strongly correlate with higher profits, particularly evident in Q4 each year, highlighting the importance of driving sales volume for profitability.

* Q1: (Query 2)

What do you want to find out?

The query aims to identify the total profit generated and the average required accuracy for each model type, helping to understand the financial performance and quality expectations associated with different models.

Insert your query here:

SELECT

mt.ModelType,

SUM(f.Price) AS TotalProfit,

AVG(f.RequiredAcc) AS AvgRequiredAccuracy

FROM

SPAIDW2A0306.dbo.fact\_table f

JOIN

SPAIDW2A0306.dbo.model\_dimension m ON f.ModelSK = m.ModelSK

JOIN

SPAIDW2A0306.dbo.modeltype\_dimension mt ON m.ModelCode = mt.ModelCode

GROUP BY

mt.ModelType

ORDER BY

TotalProfit DESC;

Insert your results here:

A screenshot of a graph

Description automatically generated

Explain what you find based on the results:

 **Top Performing Model**: **Neural Network** models are the top profit contributors, generating $633,131, indicating their popularity or effectiveness.

 **Accuracy and Profit Correlation**: **Random Forest** models, despite generating the lowest total profit, have the highest average required accuracy (65.02%), suggesting they are chosen for projects where accuracy is critical.

 **Even Distribution Among Other Models**: Profits and required accuracy are relatively evenly distributed among models like SVM, Decision Tree, and Linear Regression, highlighting a diversified use of different models with similar performance levels.

* Q2: (Query 1)

What do you want to find out?

Who the top customers are based on revenue that they bring to the Company, and approximately how much value each transaction with them is worth, finding out which customer is the most valuable.

Insert your query here:

SELECT

c.CustomerID,

c.FirstName,

c.LastName,

COUNT(f.OrderSK) AS TotalOrders,

SUM(f.Price) AS TotalSales,

SUM(f.Price) / COUNT(f.OrderSK) AS AverageRevenuePerOrder

FROM

fact\_table f

JOIN

customer\_dimension c ON f.CustomerSK = c.CustomerSK

GROUP BY

c.CustomerID, c.FirstName, c.LastName

ORDER BY

TotalSales DESC;

Insert your results here:

A screenshot of a computer

Description automatically generated

Explain what you find based on the results:

1. **Identification of Top Customers:**

The query lists customers in descending order based on sales.

* Top customers identified:
* James Hernandez
* Alexander Perez
* Amanda Washington

1. **Average Revenue Per Order:**

* The query shows the average revenue per order for each customer.
* Helps in determining if a customer provides a higher margin.
* Useful for deciding which customers to focus on for future marketing efforts.
* Q2: (Query 2)

What do you want to find out?

When do the volume of orders increase, whether revenue generated is strongly correlated with order volume and how order numbers and revenue has changed over the years by quarter.

Insert your query here:

SELECT

t.Year,

t.Quarter,

COUNT(f.OrderSK) AS TotalOrders,

SUM(f.Price) AS TotalSales

FROM

fact\_table f

JOIN

time\_dimension t ON f.TimeSK = t.TimeSK

GROUP BY

t.Year, t.Quarter

ORDER BY

TotalOrders DESC;

Insert your results here:

A screenshot of a table

Description automatically generated

Explain what you find based on the results:

1. **Highest Order Volume in Q4:**

* Total orders are consistently highest in the fourth quarter.

1. **Year-on-Year Order Growth:**

* Orders have been increasing year on year.

1. **Best Year for Sales in 2023:**

* 2023 is the best year in terms of total sales.
* Total sales in every quarter of 2023 are around double or more than double the sales in the same quarter the previous year.

1. **Revenue Growth:**

* The company has been experiencing great revenue growth.

1. **Potential Cash Flow Concerns:**

* The disparity between quarters may be a concern due to potential cash flow issues for the company.
* Q3 (Query 1):

What do you want to find out?

I want to analyze employee performance by examining the total number of orders each employee has processed and the total revenue they have generated.

Insert your query here:

SELECT e.employeesk, e.firstname, e.lastname,

COUNT(f.ordersk) AS total\_orders,

SUM(f.price) AS total\_revenue

FROM employee\_dimension e

JOIN fact\_table f ON e.employeesk = f.employeesk

GROUP BY e.employeesk, e.firstname, e.lastname

ORDER BY total\_revenue DESC;

Insert your results here:

A screenshot of a data

Description automatically generated

Explain what you find based on the results:

1. **Top Performers by Total Orders:**

* **Mia Lewis leads significantly with 651 orders processed.**
* **Emily Scott (246 orders) and Madison Morgan (241 orders) are the next highest though they have significantly fewer orders compared to Mia.**

**2. Top Performers by Total Revenue:**

* **Mia Lewis again stands out with the highest total revenue of 295,219.**
* **Emily Scott and Madison Morgan follow with 87,203 and 85,110 in revenue respectively.**

**3. Consistency in Performance:**

* **There is a strong correlation between the number of orders processed and the total revenue generated.**
* **Mia Lewis is the top performer in both orders processed and revenue generated, indicating exceptional efficiency and effectiveness.**
* Q3 (Query 2):

What do you want to find out?

I want to analyze the distribution of revenue and orders by gender, including the average revenue per order.

Insert your query here:

SELECT e.gender,

COUNT(f.ordersk) AS total\_orders,

SUM(f.price) AS total\_revenue,

AVG(f.price) AS average\_revenue\_per\_order

FROM employee\_dimension e

JOIN fact\_table f ON e.employeesk = f.employeesk

GROUP BY e.gender;

Insert your results here:

A screenshot of a computer

Description automatically generated

Explain what you find based on the results:

**1.Orders Processed:**

* **Females have processed significantly more orders (2,122) compared to males (1,321), indicating a higher workload or higher efficiency among female employees.**

**2. Revenue Generated:**

* **The total revenue generated by female employees (851,752) is substantially higher than that generated by male employees (512,085), suggesting female employees are contributing more to the overall revenue.**

**3.Average Revenue per Order:**

* **Female employees have a slightly higher average revenue per order (401) compared to male employees (387), indicating that on average, orders handled by female employees tend to generate slightly more revenue.**