ISM 545 Capstone Project

Students: Patnamshetty Rohith (rp937@nau.edu), Sai Nikhil Edulakanti (se722@nau.edu)

Part 1: Data Selection and Plan

1) We are using a Sales dataset which has categories, sub-categories, Quantity sold, Order date, shipped date, Net sales. Using these metrics, we pulled out some insights.

2) We used a fact table and three-dimension tables where we declared primary key in each of the table. Using the metrics in the tables, we have interlinked and created relationships between these tables.

❖ Fact Table:

We have taken Category as grain level for the fact table where constituting the category with other metrics.

Fact_Table

Field Name	Data Type
Sales ID	Short Text
Date Key	Short Text
Order Date	Short Text
Order ID	Short Text
Customer ID	Short Text
Product ID	Short Text
Category	Short Text
Sub-Category	Short Text
Quantity	Short Text
Product Name	Short Text
Net Sales	Short Text

❖ Dimension Table:

In our project, we have considered 3-dimension tables such as Order Dimension, Customer Dimension, and Date Dimension.

Following are the Dimension tables:

Order_Dimension

Field Name	Data Type
Order ID	Number
Order Date	Date/Time
Ship Date	Date/Time
Ship Mode	Short Text
Season	Short Text

Customer_Dimension

Field Name	Data Type
Customer ID	Short Text
Customer Name	Short Text
City	Short Text
State	Short Text
Country	Short Text
Postal Code	Number
Region	Short Text

Date_Dimension

Field Name	Data Type
Date Key	Short Text
Date	Short Text
Month	Short Text
Day	Short Text
Year	Short Text
Month Name	Short Text
Day of the Week	Short Text
Quarter	Short Text
Weekend/Weekday	Short Text
WeekNumber	Short Text
YearMonth	Short Text
YearQuarter	Short Text
YearWeekNumber	Short Text
YearWeekend/Weekday	Short Text
YearDayoftheWeek	Short Text

Part 2: Creation of the Data Warehouse

We have created our Data warehouse in Access app.

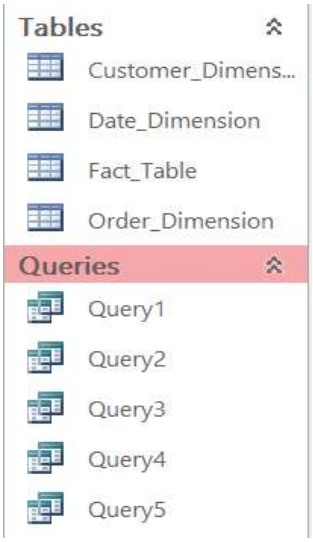
Step 1: We have exported our Excel files (Fact table, Customer dimension table, Order dimension table, Date dimension table) using External tab feature and using design view option we have defined the field names and types which exactly matches the type in excel.

Step 2: Above step will allow us to upload excel sheet with all the metrics as we have already defined the type of it.

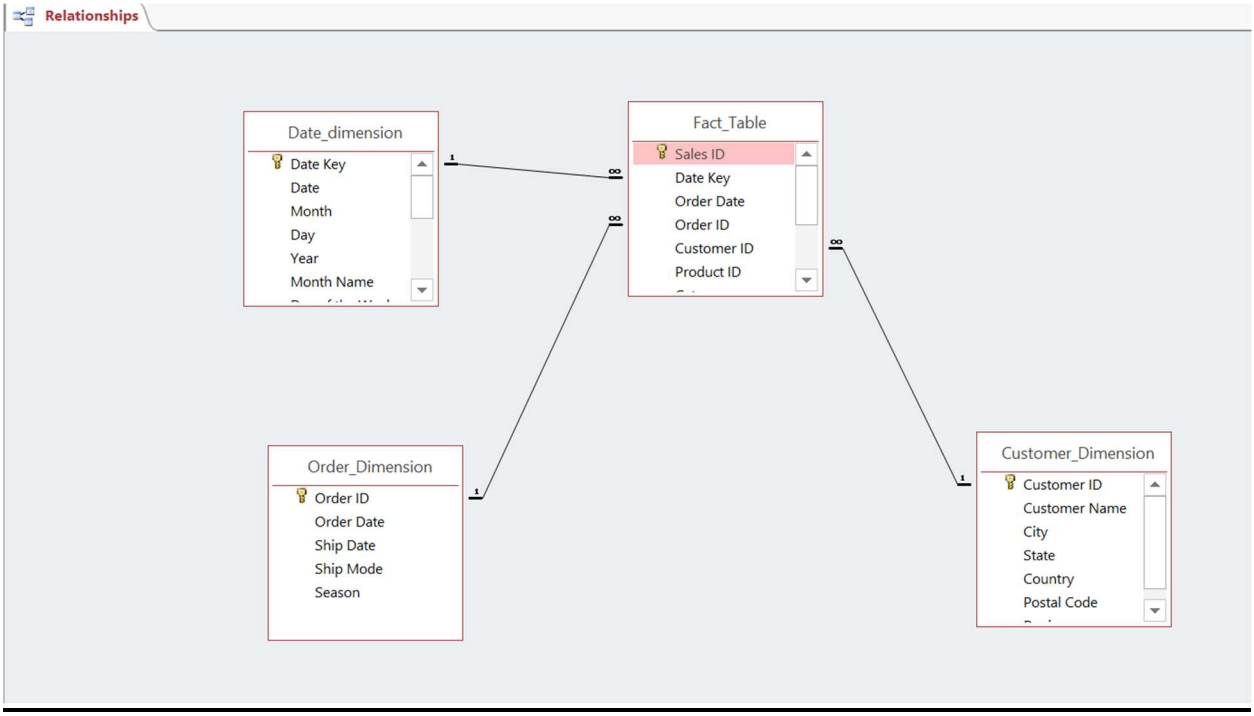
Step 3: After successfully importing the excel files, we have defined the primary key for each table in order to make relations between the tables.

Step 4: After assigning the keys to table, now it is time to create query. Using Query Design, we have created 5 queries with different insight in each.

- **Query 1**: This Query says Net sales of Category (Furniture, Office Supplies, Technology) for the years 2016-2019.
- **Query 2**: This Query gives the Season wise sales.
- **Query 3**: This Query gives the Sum of Net Sales for Sub-category's (Accessories, Appliances, Art, Binders, Bookcases, Chairs, Copiers, Envelops, Fasteners, Furnishings, Labels, Machines, Paper, Phones, Storage, Supplies, Tables) for Every month.
- **Query 4**: This Query gives the Quarterly Net sales.
- **Query 5**: This Query gives out the total quantity sold and the net sales in every state of USA for the year 2016.



The below image is the relationships we have established with a fact table and the three-dimension tables.



- **Relation 1**: This is a relation between the Fact table and the Date Dimension table where the key is Date.
- **Relation 2**: This is a relation between the Fact table and the Order Dimension where the Order ID is the key.
- **Relation 3**: This is a relation between the Fact table and the Customer Dimension where the Customer ID is the key.

Part 3: Lessons Learnt Report

This Project has taught us a lot on, how to build a Data Warehouse, how to filter excel files, how to create metrics in excel dataset using excel formulas, and many more.

There were many difficulties in doing this project.

1. Our First task was to choose a dataset from online sources. We have used Kaggle to retrieve our dataset. We have been through many datasets and finally concluded with one dataset i.e., "USA sales dataset". Sales dataset is something which has all the possible metrics necessary for a data warehousing project. We felt that this dataset was very perfect for our project as it was covering all the insights we assumed and expected.

2. After getting the dataset, it was now a task to remove the duplicates, nulls, and add dimensions required for the project. This took us a lot of time as the dataset was very improper and not structured properly. There were many duplicate data which has to be removed. We have used filters and techniques in excel and removed most of the duplicates. Now we have encountered another issue which is to remove null values. There were few blank cells in between for the date column, order date column, and quantity column. There were almost 9800 records it was little tough. We have to fill null values so that it won't affect the other metrics. The main task was to update the order date values, as it is dependent on ship date. Many of the order date values were missing and we have to rectify. The order date must be two days before to shipping date, we made sure of checking the dates and every order date is two days before of shipping date. After removing all the duplicates, null and other redundant data, we now have to add few more metrics to the dataset likely, Order ID, Season, Quarter. We have used Excel formulas to extract. We have used, "Randbetween" function to generate a eight digit Order ID, then we have used VLOOKUP formula to get Season and Quarter.

3. After having the proper dataset with all the metrics, we need, we have exported the data to a new file as Values. Now that we have the readily used dataset, it was time to create the fact table and dimension tables. This was a real task for us. As mentioned in the project, we have to create one Fact table and more than 2-dimension tables. Our tables are as follows:

    - Fact table
    - Date Dimension table, Customer Dimension Table, and Order dimension table.

4. Our Fact table had all the category related metrics namely Sales ID, Date key, Order date, Product ID, Category, Sub-category, Quantity, and Product name. It was easy to extract all of these as these were already present in Raw Dataset.

5. Our Date Dimension table was little tricky and tough. It should be having many of the metrics which we wanted in our project to make insights. So, we have started building the metrics namely, Month, Day, Year, Day of the week, Weekend/weekday, Week number, Year-Month, Year-Quarter, Year-Weeknumber, Year-Weekend/Weekday, Year-DayoftheWeek. We have used the Autofill technique to extract the Month Day and Year from the date column. We have used IF-OR formula to extract Weekend/Weekday as per day. Rest all the metrics we have extracted using CONCAT formula.

6. Next was Customer Dimension Table. It contained all the metrics related to customer namely, Customer ID, Customer Name, City, State, Country, Postal Code, and Region. All these were easy to extract as they were already present in Raw Dataset.

7. The last table was Order Dimension Table. It consists of al the metrics related to order namely, Order ID, Ship date, Order date, Ship Mode, and Season. Even these were easy to extract as they were already present in Raw Dataset.

8. After successfully creating all the Tables required for the project, we have to now upload it on Access platform to pull out the insights. There were many issues while doing this entire process. The tough task was to import the excel sheets and match with field names. There were many errors encountered during this. The Field names wouldn't match with excel and we had to again make changes in excel. This took us a lot of time as there were many metrics to satisfy. After making changes, we had to delete the previous file uploaded in excel and upload new one which has the updates metrics. Major difficulty was matching with types of the metrics. The date column should have a dimension date/time. But access didn't allow us with that type and was always pointing us errors. We then had to reach professor for help. We tried solving many of our major issues with professor help. As our professor was very kind, he always gave his best suggestions in helping us. There was one error namely "Subscript out of range" which literally made us feel afraid about the project. This error popped up like more than 50+ times. Finally, with our efforts and help from professor, we successfully imported all the excel files without any errors.

9. Initially, we related the fact table with only one-dimension table which was Sales. Later, we established two more relations with fact table. The lesson learnt here is that any model should have multiple dimension which connects multiple tables. This is necessary as we can pull insights form different dimensions based on the requirement of the user. While establishing these relations we encountered a problem where the data was not being displayed, this was due the difference in data types of the fact and the dimension table. Later, we realized the issue and changed the data types in both tables and then it ran smoothly.

10. Now it is just pulling out metrics from the tables we have updated. As we had an idea on what data to pull, we have created five queries. Each query has its own insight and meaning. Every query gave some information relating to the tables.

11. Overall, it was a great experience doing this project. This gave us a challenge to our brain and made us trick in every possibility. With all the lessons we learnt and all the labs we have been, it gave us a new experience and encouragement in completing this project.

❖ In which way this project could have been better if could repeat this assignment?

- Initially, during the labs when we worked on different dataset which professor had given us, it was quite easy as it was already structured dataset. We didn't have much to do as every instruction was given on how to perform the actions. This made us feel easy while performing the labs. Later, when we started working on the Capstone project, it made us little difficulty in performing many of the tasks. We had to do everything from scratch and had to refer the assignment which we performed during labs.

- I personally feel that; this assignment could be been better if we had another lab session on different dataset. As it was only on one dataset, we felt little difficult. If we had worked on 2-3 datasets, it would have given us more clarity and problem-solving skills on current capstone project. More practice on labs with different datasets could help us more in making much more insights and assignment would have completed faster than the time it took. I would have known what errors it encounters even before doing the project and would have made sure in not committing them.