|  |  |
| --- | --- |
|  | |
| Working With Data–Assessment 2  TU060 : Data Warehouse Modelling / Data Analysis / Machine Learning using SQL | |
|  | |
| Ciaran Finnegan – Part Time – First Year  Student No : D21124026  5/1/2022 |  |
|  |  |

Table of Contents

1 Project Overview 3

1.1 High Level Description 3

1.2 Environment Assumptions 3

1.3 Project Execution Instructions 3

2 Section A: Business Drivers 4

2.1 Background 4

2.2 Subject Area for Analysis 4

2.3 Vision and Goals for the Data Warehouse 6

2.4 Key Stakeholders 6

3 Section A: Data Modelling 7

3.1 Data Warehouse Schema 7

3.2 Reasons for Design 8

4 Section A: Implementing the Data Warehouse 14

4.1 Implementation using SQL Scripts 14

5 Section B: Data Analysis Using SQL 15

5.1 Data Analysis Objectives 15

5.2 Data Analysis – Context Data 16

5.3 Data Analysis – KPIs 17

6 Section C: Machine Learning Using SQL 19

6.1 Overview of ML Process 19

6.2 Creating and Populating Fact Table for ML Modelling 19

6.3 Preparing Training and Test Data Sets 19

6.4 Create VIEW for Predicted Values 19

6.5 Evaluating Models 19

7 Appendices 20

7.1 Appendix 1 – SQL Scripts to build the Data Warehouse 20

7.2 Appendix 2 – SQL Scripts to Populate Data Warehouse Dimensions 20

7.3 Appendix 3 – SQL Scripts To Populate Data Warehouse Fact Table 20

7.4 Appendix 4 – SQL Scripts For All SQL Queries 20

7.5 Appendix 5 – SQL Scripts For ML Process 20

8 References 21

8.1 Data Warehouse Design 21

# Project Overview

## High Level Description

This document covers the design, implementation and observations on all parts of the January 2022 CA(2) for the working With Data module in the TU060 Part Time/first Year MSc in Science (Data Science) course.

## Environment Assumptions

The ..

## Project Execution Instructions

The project data warhouse solution is executed from the

# Section A: Business Drivers

## Background

The ...

## Subject Area for Analysis

In this assignment I have chosen to focus on an analysis of **Sales performance**…?

Determining which type of ? activity ..

### KPIs

To start with definitions;

A **Key Performance Question** (KPQ) is a management question that captures what a manager needs to know in order to better understand the performance of the company or organisation.

A **Key Performance Indicator** (KPI) is a measure to provide managers with the most important performance information required to enable them, or their stakeholders, understand the performance of the business.

For this assignment, I have identified the following KPIs to assist management in understanding Sales performance;

**Customer KPI**

..

**Financial Performance KPIs**

..

**Employee KPIs**

..

### Action Plan

This project will detail the steps to create the Data Warehouse and then follow through on the techniques, and applications to analyse the data stored in this Data Warehouse.

In a real world scenario, Senior Management, and/or other stakeholders, would determine an action plan, based on dashboards and KPI reports, to improve Sales performance.

## Vision and Goals for the Data Warehouse

The Data Warehouse created in this project is...

## Key Stakeholders

The KPI reports produced in this project are Sales performance data based on quarterly, half-yearly, and yearly timeframes.

This information is thus not expected to be updated daily and is more strategic in value.

The key stakeholder for the type of reporting produced in this project would therefore be;

* **Senior Management.** Sales performance data that could change product lines or seller rewards is information that can only really be actioned by those within Our telecomunications company.. who control major resourcing and policy decisions.
* **Employees.** Those in the salesforce itself in ...our telecomunications company.. The dashboards provide a means for sellers to assess their performance when judged against the performance of their peers. It also allows for sellers to see which product lines are working best for them. The information should assist in either providing incentives or information for their own personal business decisions.
* **Company owners.** Is the Our telecomunications company.. company profitable? Are there KPIs that show the business is being well run? If reports show that sales are consistently growing over time then this allows the owners of the company to have confidence in Senior Management. If sales were falling then the owners of ...our telecomunications company.. may look to take corrective action and seek new management.
* **Investors.** Is ...our telecomunications company.. a company that appears to have a promising future? Is it a candidate for external investment? The dashboard I generate for 1997 Sales performance contains the type of reports that would feature in a year-end report to market investors.

# Section A: Data Modelling

## Data Warehouse Schema

The Data Warehouse for this project will be implemented with a Star Schema design.

This involves one central Fact table surrounded by a number of Dimension tables.

<Star Schema Diagram>

The operation database from which this Data Warehouse is build has a normalised relational structure.

To optimise queries on the Data Warehouse the tables are effectively ‘de-normalised’.

The Dimension tables contain descriptive information. The Fact table contains keys to all the dimension table Primary keys, and all the measurable attributes required to meet the reporting purpose of this Data Warehouse.

## Reasons for Design

The Data Warehouse in this project is built following the design principles as described in Kimball’s four step process.

1. Identify the Business Process. Do not re-model the Business Department / Area.
2. Identify the Grain.
3. Choose the Dimensions.
4. Choose the Facts

These steps will be applied to the creation of a new Data Warehouse for this project, but this process could also be applied to the enhancement of an existing Data Warehouse to include a new business process reporting objective.

The objectives in creating the star schema model for the Data Warehouse are;

* Be simple.
* Be easy to use.
* The ETL process loading into these tables should be as simple as possible.
* Queries should perform well with SQL, or other Business Intelligence tools (which are not part of this assignment task).

### Identify the Business Process

This is the first step in designing the Data Warehouse.

A ‘Business Process’ can be defined as a natural operational activity performed in the organisation, in this case Our telecomunications company.., that is supported by some form of data collection.

The following should be considered when identifying the process on which we wish to focus;

1. **Look at the business process not the business department.** This allows for data to be collated and reported on in a more consistent manner across the organisation. It helps in avoiding duplication of data, which might occur if we replicate the structure of business units in the Data Warehouse. In this project we are looking at the Sales process by measuring Sales performance and we are not looking to build a Northwind Sales Department report. I will focus on the Sales business process to provide a singular vision of Sales performance in Our telecomunications company.., which actually somewhat spans across business departments.
2. **Assess impact and risk in reporting on the chosen business process**. Impact is generating reports that the business actually want on a regular basis. I am assuming that Our telecomunications company.. are dependent on quarterly, half yearly and yearly reports on Sales performance to make decisions on optimising its sales force and product range. The risk with reporting on some business processes for Our telecomunications company.. is that there are some data gaps in the database and this could complicate the reporting in certain areas. For example, there are NULL values in the ‘Region’ column in the Employees table in the Our telecomunications company.. operational database. This could be patched but it is additional effort and complexity. (Arguably it would not be very difficult to work around but we can avoid the risk by choosing a business process that is not reliant on that column).
3. **In a real world scenario, the business users would provide guidance on my impactful business process for them.** Business users can also help decipher complex business processes. The Our telecomunications company.. business model is relatively simple to understand, so it is my assumption the Sales performance is an important business area for the company.

From the above points, it is important to emphasise that the Data Warehouse I am creating is intended to service reporting needs for the entire Our telecomunications company.. organisation.

### Identify the Grain

This is the second step in designing the Data Warehouse.

This is the most important phase of the design process. Redesigning a Data Warehouse at a later date to increase the level of granularity could be an expensive and time consuming process.

The resultant Fact table will be at the centre of our star schema. This table contains all of the measurable facts about the captured business process. We will use the Fact table to extract information on how successful, or not, the Sales performance is, and has been, in ...our telecomunications company..

I have followed three particular guidelines in my project to identify the correct level of granularity when considering the design of this Data Warehouse.

1. **What is represented by one ‘Fact’ row?** What level of granularity is captured?
2. **Choose the most atomic level of information.** The data cannot be meaningfully subdivided any further. It also allows for easy and effective aggregations.
3. **Allow scope for future reporting requirements**. It is hard to predict future user requirements so the granularity is important to allow further, possibly ad-hoc, reporting requirements.

In my Fact table the focus is on measuring <what?>. Thus in my dimensional model one Fact row represents *one product sold to a customer by one seller on one given day*.

This is a lower grain than what a seller sold in a day, as the seller may make multiple sales on one day.

My SQL queries, as shown in Section 5 of this document, look at <what?> performance at a quarterly, half yearly, and yearly basis. However, it is still good practice to store information in the Data Warehouse at a daily level to future proof for other Sales performance reporting requirements.

### Choose the Dimensions

This is the third step in designing the Data Warehouse.

Guidelines for this process can be summarised as follows;

1. Who, what, where, when?
2. Best attributes are descriptive.
3. De-normalizes design focuses on high performance reads.
4. Use smallest data types possible.

To capture the attributes of the Sales performance process in Our telecomunications company.., I need to have the information on **who** (Seller) sold which **what** product to **whom** (Customer) and **when**. Fields like phone numbers, and postal codes are not used in any of my KPIs and hence are not included in any of the dimension tables.

This question dictated the choice of the dimension tables I selected for my Data Warehouse schema;

* Product (*dw\_dimtblProduct*)
* Customer (*dw\_dimtblCustomer*)
* Employee (*dw\_dimtblCustomer*)

As an example, the choice of supplier is not relevant for reporting on the Sales performance process, so I did not include a Supplier dimension table.

The TimeDate dimension table (*dw\_dimtbltimeDate*) is built to provide additional date granularity and a conversion of the date into an integer format to improve reporting performance.

A new ‘surrogate key’ has been created for each of the dimension tables. It is a simple numeric value that I have set to increment in the SQL scripts used in the CREATE TABLE routines.

The surrogate key is necessary to uniquely identify each row in the dimension table and to avoid any confusion with the source Primary Keys from the Northwind operational database (‘Northwind’). This is particularly useful if the key structure in the Our telecomunications company.. operational database changed in the future. Such changes will not then have a knock on impact on the Data Warehouse and reporting applications should still be valid.

The surrogate keys of each dimension table are usually simple integer values and are also added to the Fact table. This is done to minimise the number of joins needed to fetch data, which improves the response time of queries (as does the use of simple integer key values).

### Choose the Facts

This is the fourth step in designing the Data Warehouse.

The Fact table exists at the centre of the star schema, as can be seen in Section 3.1.

Defining the measures for the Fact table should follow guidelines such as these;

1. **How does the business measure success?** For Sales performance we are looking at the Net Sales totals and what Products generate the most, who is the best employee at selling based on sales, and where are the highest sales being made (based on the Net Sales amount)?
2. **The best measures are fully additive**. It should be possible to roll up the measures and easily perform aggregations. In my Tableau workbook I generate reports that display Net Amount values totalled by geography and individual sellers (as an example).
3. **Data access tools, such as Tableau, are suitable for non-additive measures.** Year To Date averages are calculated in one of my dashboards, but would not be a meaningful unit of data in the Fact table.

The Net Sales amount is a relatively simple metric to report on, as it is stored in the Fact table.

I have also included the GrossTotal amount, OrderQuantity, DiscountTotal, and DaysToShip in the Fact table. These measures are ultimately not used in the Tableau dashboards in Section 5 but are included them to allow for possible future reports.

The Net Sales amounts stored at the level I have chosen in the Fact table allows for the KPIs in Sales performance to be easily reported in Tableau (although with supplementary SQL in some cases).

# Section A: Implementing the Data Warehouse

## Implementation using SQL Scripts

Code in appendix…

Some explanatory bluff here..

# Section B: Data Analysis Using SQL

## Data Analysis Objectives

The..

There is an overarching theme to this dashboard;

* **What** products sell the best?
* **Where** are the best markets?
* **Who** are the best sellers?
* **How** are we performing?

The information in the dashboard is very high level but it intended as an ‘at a glance’ overview for Senior Management. The message is clear and focuses on the key successes of 1997.

## Data Analysis – Context Data

**Tile 1: Top 5 Products**

* The ‘Top 5 Products’ tile follows a ‘Golden Rectangle’ dimension as it is approximately 50% wider than tall. That dimension suits the graphic as is allows for an easy comparison of net sales amount per product.

**Tile 2: Net Sales By Country**

* The ‘Net Sales By Country’ tile uses a graduated scale of blue colours to show the countries into which Our telecomunications company.. sell. The varying shades of blue show the varying quantity of net sales. This is more effective in showing a natural hierarchy of where sales are highest. It is also easier to read for someone with challenges in colour perception.

**Tile 3: Top 5 Sellers**

* The ‘Top 5 Sellers’ tile follows the clear principle of ‘don’t make me think’. It is a simple representation of top seller performance.

**Tile 4: Total Half Yearly Net Sales (1997 + 1998)**

* Using a table in the ‘Total Half Yearly Net Sales (19967 + 1997)’ tile is a more effective way to display this numerical data and it provides the background the percentage tiles below.

**Tile 5: 1997 H1 Growth (v. 1996 H2)**

* Percentage values compliment other graphical and tabular data in this spreadsheet. A ‘+’ has been added to emphasis the performance success. The font in the percentage tile has been increased to draw user attention

## Data Analysis – KPIs

In Section 5.1of this document the …

### KPIs: Topic One Data Analysis

The purpose of these queries..

**KPI 1: The Top 5 Products**

This metric informs the Sales team about what products are selling well, and into which category they fall.

It could help Senior Management assess where future production efforts can be diverted to maximise profits.

**KPI 2: Net Sales By Country**

Where are we selling the most products? This graphic shows where Net Sales were highest in 1997.

### KPIs: Topic Two Data Analysis

The purpose of these queries..

**KPI 1: YTD Net Sales (1998)**

This metric informs the Sales team about who is performing best amongst the sellers in the H1 1998 period (Q1 and Q2)

**KPI 2: Performance of each seller against the yearly Net Sales average (historical)**

This is a more complex set of KPIs and provides a view across the entire available timeline in the Our telecomunications company.. Data Warehouse of how sellers are performing compared to their peer.

# Section C: Machine Learning Using SQL

## Overview of ML Process

In the..

## Creating and Populating Fact Table for ML Modelling

In the..

## Preparing Training and Test Data Sets

In the..

## Create VIEW for Predicted Values

In the..

## Evaluating Models

In the..

# Appendices

## Appendix 1 – SQL Scripts to build the Data Warehouse

In the..

## Appendix 2 – SQL Scripts to Populate Data Warehouse Dimensions

The Dimension tables in the Data Warehouse are populated..

## Appendix 3 – SQL Scripts To Populate Data Warehouse Fact Table

The Fact table in the Data Warehouse is populated in the by executing the SQL commands in the attached \*.SQL file.

## Appendix 4 – SQL Scripts For All SQL Queries

The

## Appendix 5 – SQL Scripts For ML Process

The

# References

## Data Warehouse Design

In additional to the class Module notes I followed the data warehouse design principles that were discussed in these two YouTube training videos;

***Designing Your Data Warehouse from the Ground Up* -** <https://youtu.be/patBYUGwsHE>

***Implementing a Data Warehouse with SQL Server, 01, Design and Implement Dimensions and Fact Tables* -** <https://youtu.be/StoWu2A8Ufs>