|  |  |
| --- | --- |
|  | |
| Working With Data–Assessment 2  TU060 : Data Warehouse Modelling / Data Analysis / Machine Learning using SQL | |
|  | |
| Ciaran Finnegan – Part Time – First Year  Student No : D21124026  4/1/2022 |  |
|  |  |

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 for Assignment 4

2.1 Background and Goals 4

2.2 Subject Area for Analysis 4

2.3 Key Stakeholders 4

2.4 KPIs 5

3 Section A: Data Modelling 6

3.1 Reasons for Design 6

3.2 Data Warehouse Schema 12

4 Section A: Implementing the Data Warehouse 13

4.1 Implementation using SQL Scripts 13

5 Section B: Data Analysis Using SQL 14

5.1 Data Analysis Objectives 14

5.2 Data Analysis – Context Data 15

5.3 Data Analysis – KPIs 16

6 Section C: Machine Learning Using SQL 18

6.1 Overview of ML Process 18

6.2 Creating and Populating Fact Table for ML Modelling 18

6.3 Preparing Training and Test Data Sets 18

6.4 Create VIEW for Predicted Values 18

6.5 Evaluating Models 18

7 Appendices 19

7.1 Appendix 1 – SQL Scripts to build the Data Warehouse 19

7.2 Appendix 2 – SQL Scripts to Populate Data Warehouse Dimensions 19

7.3 Appendix 3 – SQL Scripts To Populate Data Warehouse Fact Table 19

7.4 Appendix 4 – SQL Scripts For All SQL Queries 19

7.5 Appendix 5 – SQL Scripts For ML Process 19

8 References 20

8.1 Data Warehouse Design 20

# 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 was developerd in ORACLE SQL Developer.

All SQL scripts used in the assignment have been embedded in the document but also submitted separately as ***\*.SQL*** files.

These SQL files are grouped in ***\*.ZIP*** files that are numbers to indicate the sequence in which they should be executed in a test ORACLE database.

## Project Execution Instructions

The ***\*.ZIP*** files should be unpacked in the order of the file names.

That is, *<filename\_****1****>.zip* is unpacked first and then excuted to update the target ORACEL database, before any action is taken with *<filename\_2>.zip,* and so on.

The CREATE TABLE SQL Scripts creates the source data table with specific table names that are then referenced by later SQL Scritps.

# Section A: Business Drivers for Assignment

## Background and Goals

A new data warehouse model is being designed and built to improve customer profile data for a this telecommunications company.

This report will explain to the key stakeholder group;

* The primary objectives driving the structure of the data warehouse.
* The range of reporting data that will be available from the warehouse.
* The predictive analysis that will be available to stakeholders in terms of possible customer churn.

## Subject Area for Analysis

In this assignment I have chosen to focus on an analysis of **Revenue performance.**

Which customers generate the most financial value for company based on their activity, profile, and call plans?.

## Key Stakeholders

The KPI reports produced in this project are Revenue performance data based on ? 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.** Revenue performance data that could alter contract plans or rate types. These are the types of decisions that can only be actioned by those within our telecomunications company who control major resourcing and policy decisions.
* **Customer Services agents.** Those in the salesforce in our telecomunications company. The SQL queries in this assignment can be the basis for dashboards that provide a means for Customer Service to focus their attention on customers generating higher revenues.
* **Company owners/potenail investors.** How much revenue is our telecommunications company generating?

## 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 Services KPIs**

..

**Customer Revenue KPIs**

..

# Section A: Data Modelling

## Reasons for Design

The Telecomunications 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.
* Any 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 for 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 Revenue process by measuring Revenue performance but we are not looking to build a Customer Service Department report. This assignment will look at Customer revenue generating activity across the company.
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. The assignment will focus on identifying those types of customer who generate the most revenue, and being proactive to keep them in the business.
3. **In a real world scenario, the business users would provide guidance on a data warehouse procees.** Business users can also help decipher complex business processes. In this assignment we already have conduced an engagemenet with key stakeholder, which will feed into the structure of tables and the SQL queries/reports that will be executed.

For this assignment, the business process is to capture revenue generating activity by the customers of the company.

Building the data warehouse will allow the company to have a better analytical view of the revenue streams per customer, which will then help information subsequent decision making processes.

### 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 the key Revenue data points for customers in this telecommunications 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 revenue from customer calls. Thus in my dimensional model one Fact row represents *one call event to/from a specific customer at a specific time.*

This is a lower grain than call events by a customer in a given day, as the customer may make multiple calls or voicemails on one day in different time periods (peak or off-peak).

The term ‘call event’ is significant because the telecommunications company distinguishes, in separate database tables, between;

* A voice call made ***by a*** customer.
* A voicemail left ***by a*** customer.
* A call from a Customer Service agent ***to a*** customer.

We will not subdivide out the call ***to*** and ***from*** a customer. It will be assumed that calls made by a customer are a charge to them and revenue to the company. To maintain this level of granularity, calls from a Customer Services agent are a charge to the telecommunications company itself and will be represented as a *negative* revenue item.

### 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 Revenue performance process for our telecomunications company, I need to have the information on **who** (Customer) was involved in **what** call-event and **when**.

This question dictated the choice of the dimension tables I selected for my Data Warehouse schema, and the attributes in these tables;

* Time (*dw\_dimtblDateTime*)
  + Date Attr 1
  + Date Attr 2
  + …
* Customer (*dw\_dimtblCustomer*)
  + Phone Number – The Fact table stores the Connection ID, therefore this text variable can be stored in the Customer DIM table.

* + Plan Name – text description of Customer Plan.
  + …
* Call Event (*dw\_dimtblCallEvent*)
  + Connection Id – unique identifier for the call event.
  + Call event type – text description of call type – peak, roamining, voicemail, etc.
  + Attr 2
  + …

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 ‘operational’ database tables of the telecommunications company. This is particularly useful if the key structure in the 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 3.2.

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

1. **How does the business measure success?** For Revenue performance we are looking at the charge totals for call events, and which customers are generating the most revenue?
2. **The best measures are fully additive**. It should be possible to roll up the measures and easily perform aggregations. In the SQL Scripts in Section 5 I <***give example***> (as an example).
3. **Data access tools, such as Tableau, PowerBI, (or even SQL Scipts) are suitable for non-additive measures.** ***Year To Date averages***? are calculated in the output of one of my SQL Scripts, but would not be a meaningful unit of data in the Fact table.

The facts are numeric values that correspond to the grain of the table, as defined in Section 3.1.2.

The Fact table ***dw\_facttblCallRevenue*** will be created in out Telecommunications company data warehouse.

The colums for this Fact table can be identified as follows;

1. Date Time Foreign Key – link to DateTime Dimension table.
2. Customer Foreign Key - link to Customer Dimension table.
3. Call Event Foreign Key - link to Call Event Dimension table.
4. Customer Plan Id – identifier for customer contract type.
5. Call Event Type – call (peak or off-peak), voicemail, Customer Service call
6. Duration of Call Event
7. Charge Generated for Call Event

The Charge Generated (Revenue) amount per call event is a relatively simple metric to report on as it is stored in the Fact table, and allows for more straightforward, and performant, SQL queries.

## 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.

# 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>