14/12/2020

**Abdullah Bin Zubair**

[abdullahbinzubair570@gmail.com](mailto:abdullahbinzubair570@gmail.com)

[az5645a@greenwich.ac.uk](mailto:az5645a@greenwich.ac.uk)

Data Warehousing

Project Report

Contents

[Introduction 2](#_Toc69699788)

[Star Schema Design 3](#_Toc69699789)

[Importing tables into Oracle 4](#_Toc69699790)

[Step 1: 4](#_Toc69699791)

[Step 2 6](#_Toc69699792)

[Implementation of Dimensions and Fact 9](#_Toc69699793)

[DIMENSION 1: REQUEST\_DIM table 9](#_Toc69699794)

[DIMENSION 2: TYPE\_COVER\_DIM table 9](#_Toc69699795)

[DIMENSION 3: LOCUM\_DIM table 10](#_Toc69699796)

[DIMENSION 4: PRACTICE\_DIM 11](#_Toc69699797)

[DIMENSION 5: TIME\_DIM table 12](#_Toc69699798)

[FACT 1: SESSIONS\_FACT table 12](#_Toc69699799)

[Data cleaning plan and practices 13](#_Toc69699800)

[Cleaning table 1: LOCUMDETAILS 13](#_Toc69699801)

[Cleaning table 2: PRACTICEDETAILS 14](#_Toc69699802)

[Cleaning table 3: SESSION 16](#_Toc69699803)

[Database Mapping 17](#_Toc69699804)

[Data Loading 18](#_Toc69699805)

[Step 1: Using cursor to load LOCUM\_DIM 18](#_Toc69699806)

[Step 2: Using cursor to load PRACTICE\_DIM 19](#_Toc69699807)

[Step 3: Using cursor to load REQUEST\_DIM 20](#_Toc69699808)

[Step 4: Using cursor to load TYPE\_COVER\_DIM 21](#_Toc69699809)

[Step 5: Loading TIME\_DIM 22](#_Toc69699810)

[Step 6: Using cursor to load SESSION\_FACT 23](#_Toc69699811)

[Information Queries 24](#_Toc69699812)

[Summary 24](#_Toc69699813)

# Introduction

The National Health Service (NHS) gives funding each year to local agencies to provide temporary

staff to cover for doctors and general practitioners when they are either on holiday or absent from

the workplace due to illness or any other reason. When a surgery, has a vacancy they contact an agency, by phone, and request a doctor, called a locum, to cover for a given period, called a session. NHS have been given a database (MS Access) of an agency and they want to get useful information out of it but the database is not in right structure. The data inside needs to be cleaned and the schema also requires to be changed to carry out necessary transactions.

As a Datawarehouse specialists, NHS have asked us to carry out the ETL process and output the required information from the database of agency. We divided the project into 4 parts – Importing, Implementing, Cleaning and Loading of data. In this report we will demonstrate how we were able to fulfil requirements of NHS by building a Data warehouse for them. We will be using Oracle 12c for building the Data warehouse.

# Star Schema Design

After analysing the NHS requirements and Locum agency’s database, following schema was designed and implemented:

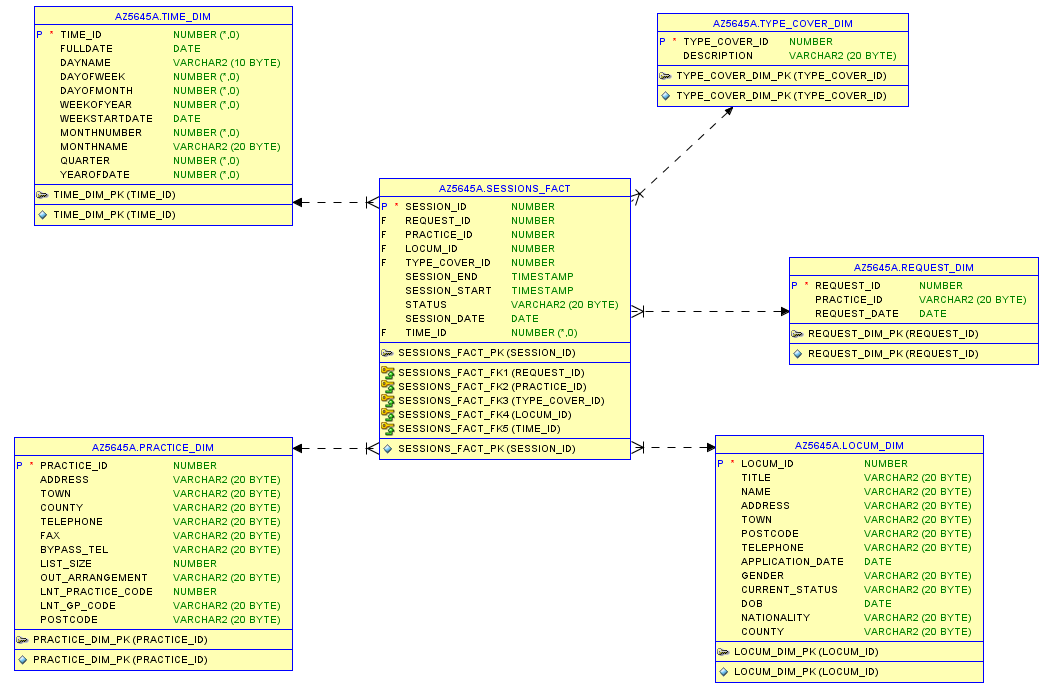


Figure 1- Star schema

This Data Warehouse schema is built just to fulfil the needs of NHS as described in coursework document. The information needed is:

* The list of Sessions filled by type of cover by month
* The number of Requests made by type of cover by week
* The number of Locum requests made by county by month
* The number of Sessions covered by month
* A list of Locum requests by town by week
* The number of requests not covered by month

In order to implement and populate the above schema, only these tables were imported and cleansed from the given agency database:

* locumDetails
* locumRequest
* practiceDetails
* session
* Template\_Type of cover

# Importing tables into Oracle

There are several ways of importing data into an Oracle schema from MS Access. These include:

* Saving the MS Access table as a flat file then importing it into Oracle using SQL\* Loader utility.
* Using Oracle’s Migration Workbench, and
* Creating a connection using the Open Database Connectivity (ODBC) utility interface.

In this project, we used the ODBC connectivity interface as it allows for maximum interoperability as the interface allows for data access in and is independent of multiple database management systems (DBMS) (Engel, et al., 2020).

The following steps are followed to import the table Template\_Type of Cover from the sample database provided for this project (Anon., 2020).

## Step 1:

Launch the ODBC connectivity utility

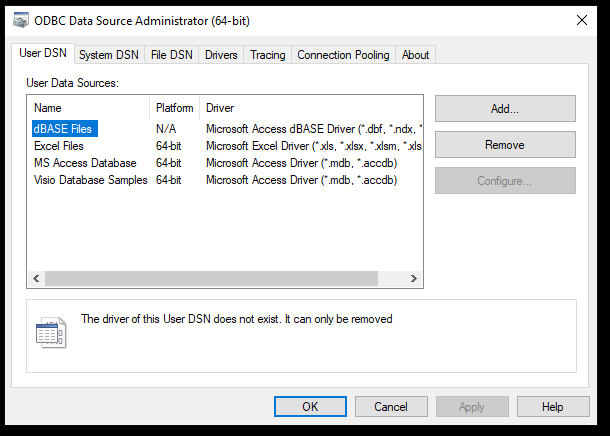


Figure 2: ODBC connectivity interface

As this is a new connection to the database, click on Add to create a new Data Source Name (DNS).

Select Oracle in Client64 and click Finish.

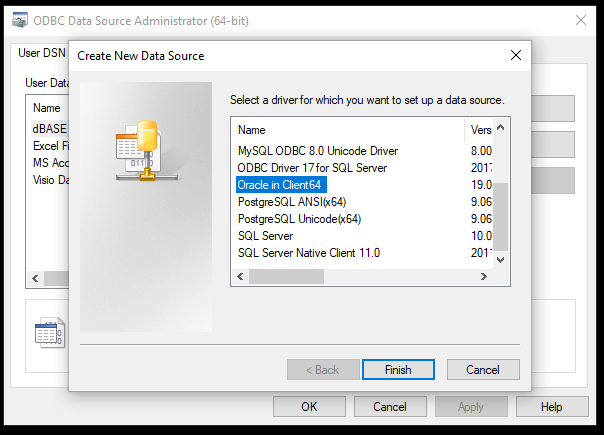


Figure 3:Selecting the Oracle in Client64 driver to set up new data source

The data source name – obiwan and TNS are entered. To test the connection, the username and password are entered as shown and validated with OK.

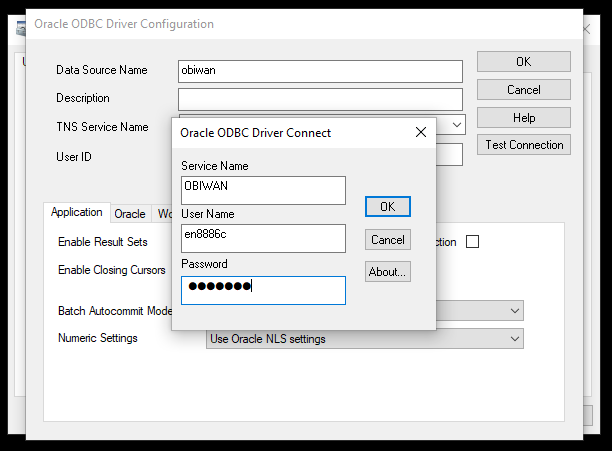


Figure 4:Naming data source and testing new connection

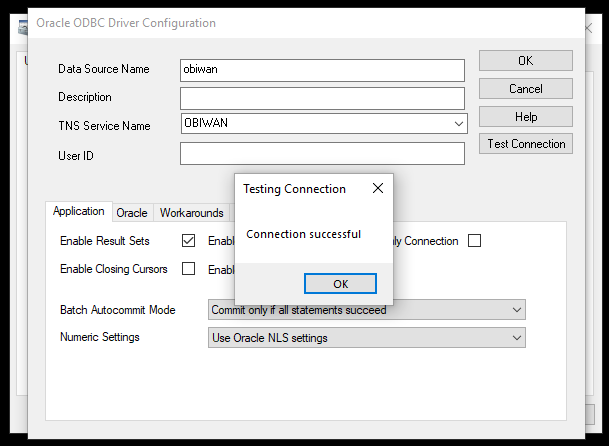


Figure 5:Testing connection successful

With the new connection validated, we are now ready to go to Step 2 to start importing the tables.

## Step 2

With the MS Access database open, right-click on the Template\_Type of Cover table. On the dropdown table, select Export and then ODBC.

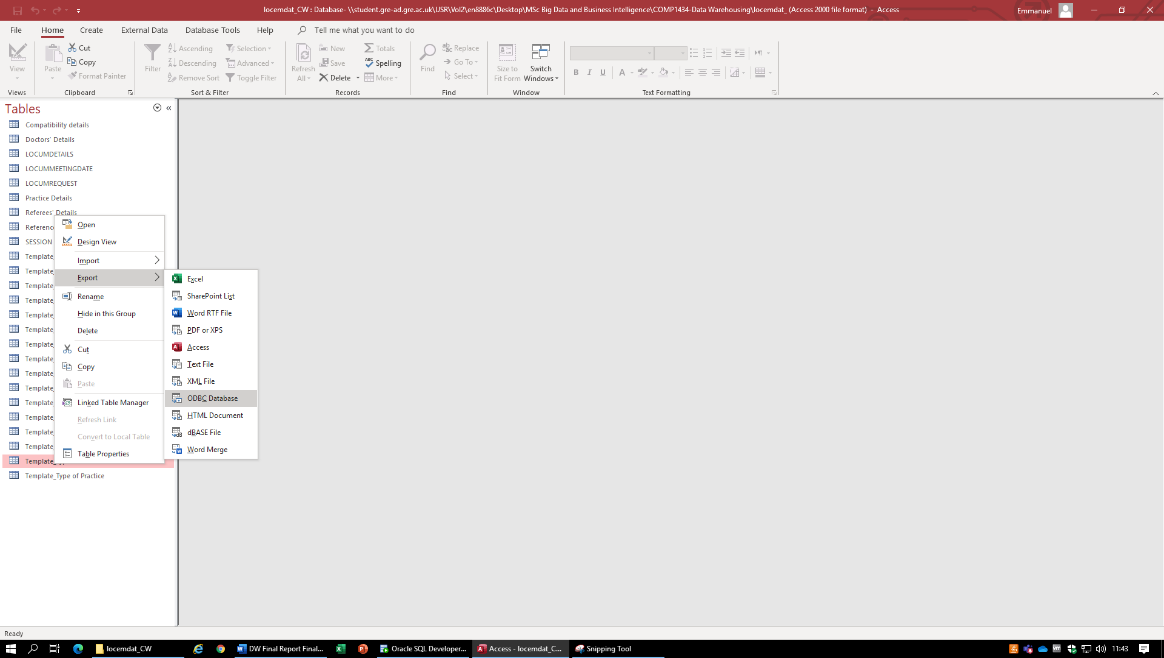


Figure 6:Selecting ODBC to connect to SQL server

With the table’s name entered, click OK.

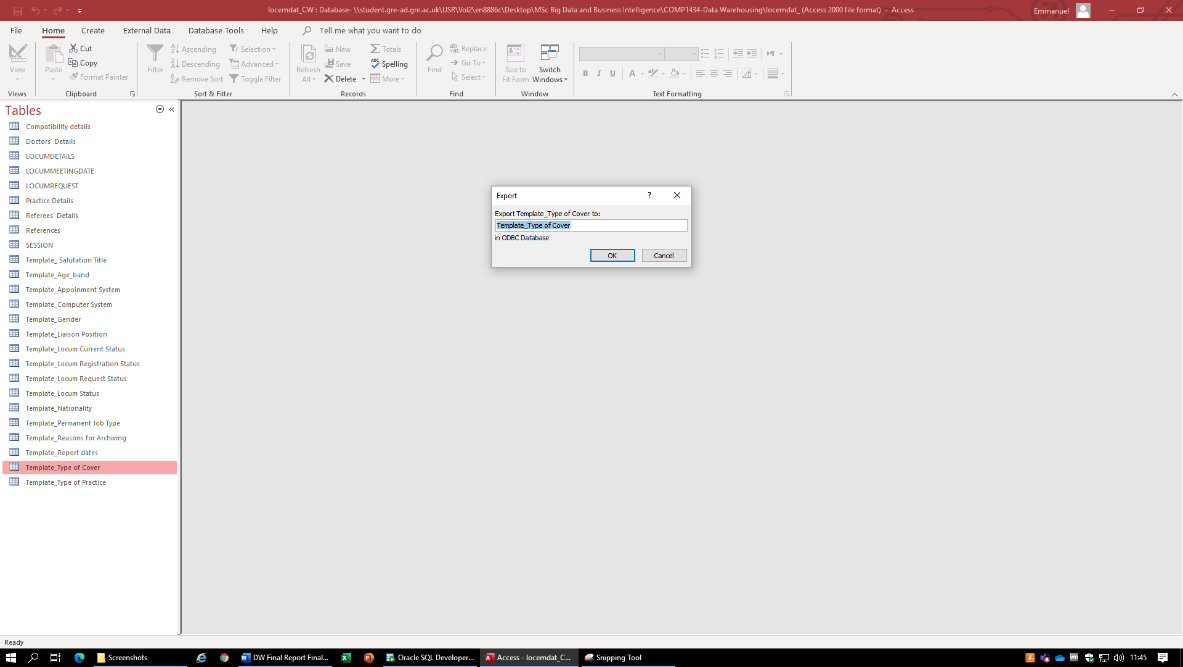


Figure 7

Now select the DNS obiwan and click OK.

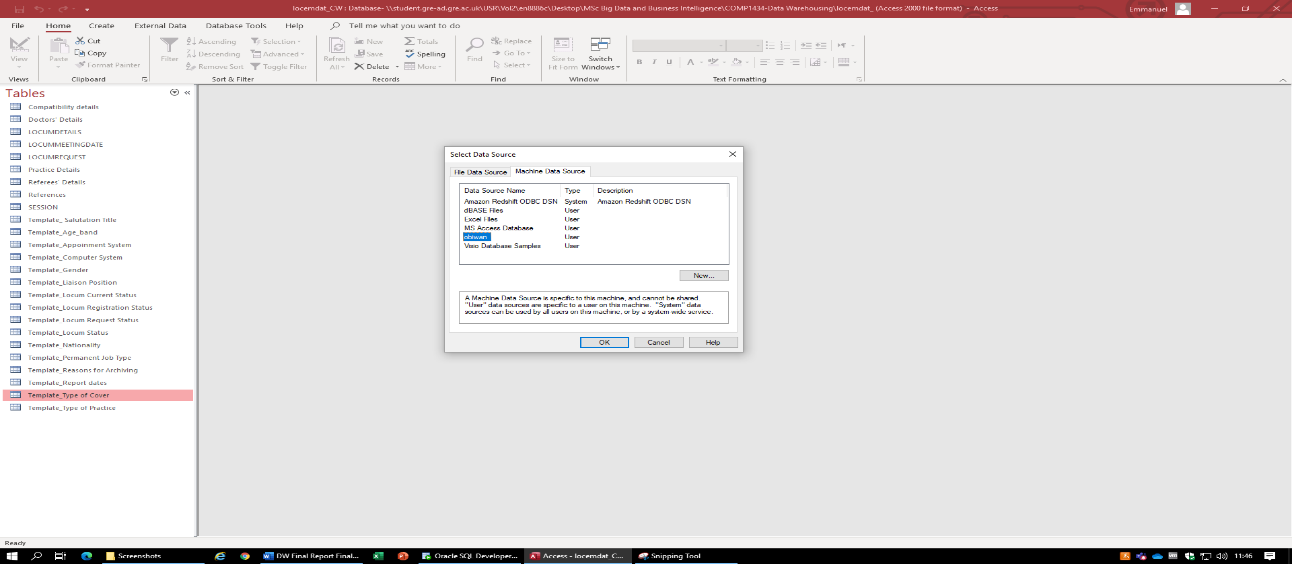


Figure 8: Selecting the data source

Now enter credentials and click OK to connect the driver.

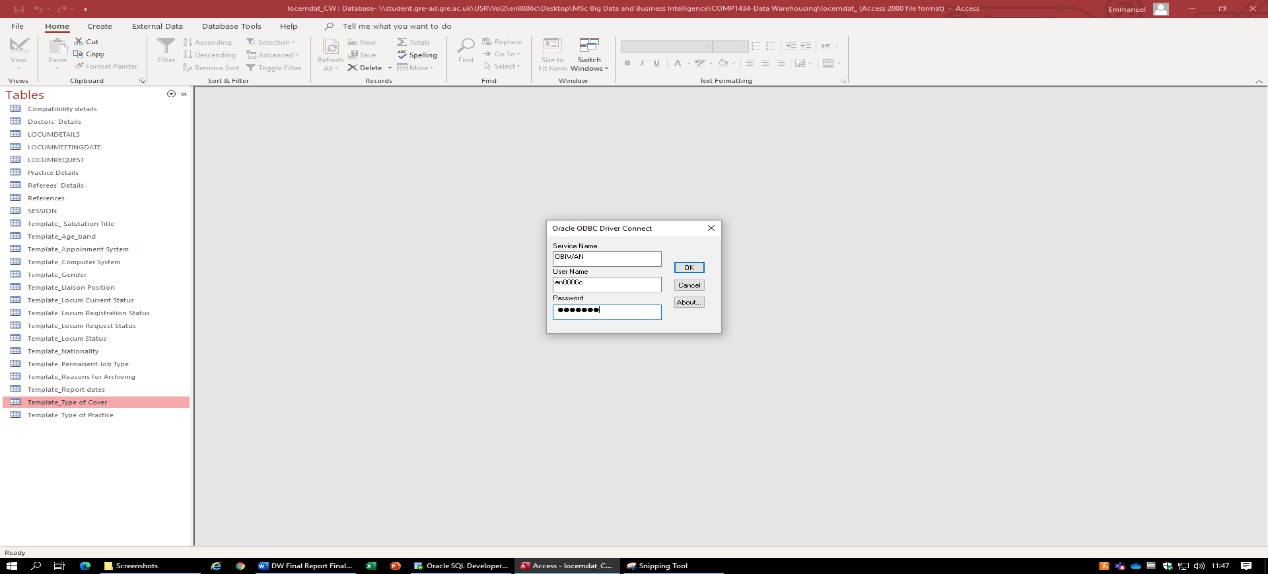


Figure 9

Table Template\_Type of Cover exported successfully. The export steps are saved to be used to export the remaining table required for the project.

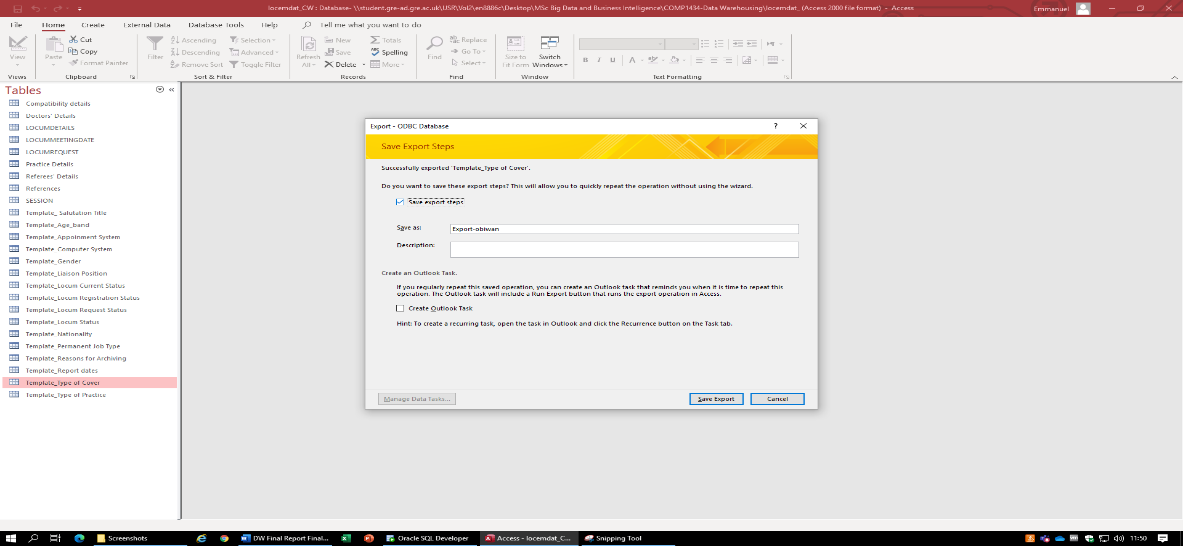


Figure 10: Export successful

The same process is repeated using the saved export steps to export the following tables:

* LOCUMDETAILS,
* LOCUMREQUEST,
* Practice Details and
* SESSION

The screenshot below shows the exported table in the staging area in oracle ready to be cleaned.

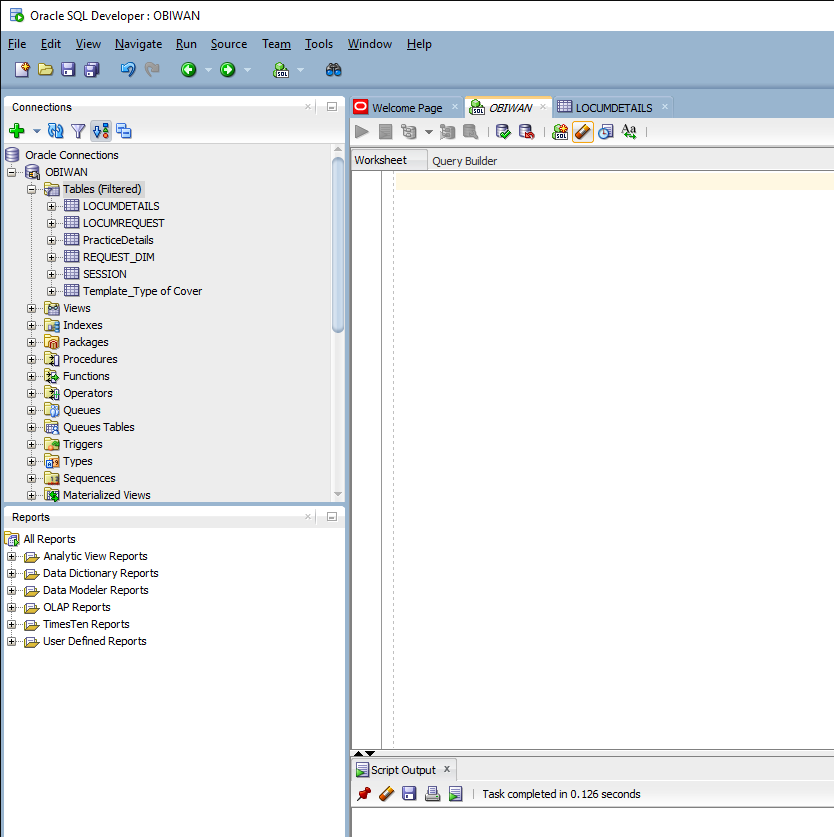


Figure 11:Exported tables in staging area in Oracle

# Implementation of Dimensions and Fact

After finalising the star schema for the project, I moved on to create the tables required:

## DIMENSION 1: REQUEST\_DIM table

This table reflects the LOCUMREQUEST table in the given database. Only the columns which made sense and were necessary to query required data were chosen. The command to create this table is:

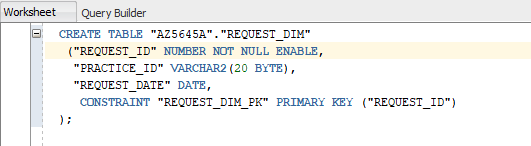


Figure 12

## DIMENSION 2: TYPE\_COVER\_DIM table

This dimension table relates to TEMPLATE\_TYPE\_OF\_COVER table in given database. It just includes an id and a description of the type. Its creation command is:

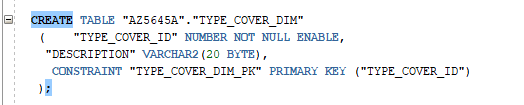


Figure 13

## DIMENSION 3: LOCUM\_DIM table

This is a very important table in this star schema and this relates to LOCUMDETAILS table in the given database. LOCUMDETAILS had many columns in it, some of them totally unnecessary so I dropped those. Columns which were taken into consideration were:

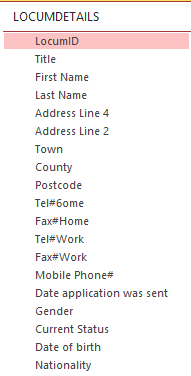


Figure 14

However, some of the columns were combined:

* FIRSTNAME and LASTNAME were joined and named NAME.
* ADDRESSLINE4 and ADDRESSLINE2 were joined as ADDRESS.

Command to create this dimension is:

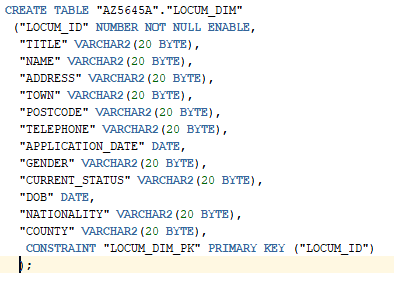


Figure 15

## DIMENSION 4: PRACTICE\_DIM

This table relates to PRACTICEDETAILS in the given database. LOCUMDETAILS had many columns and not all of them were of any use to me so I decided to pick just those columns which seemed to be logical and were necessary for the required tasks. Below are listed the columns of PRACTICEDETAILS which were taken into consideration. The highlighted columns were not considered and thus these were not added to the dimension:

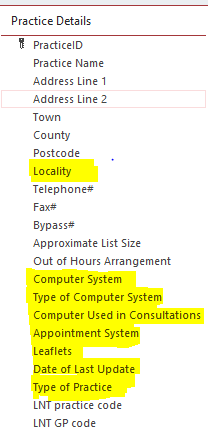


Figure 16

NOTE: Highlighted columns were not chosen to be in the PRACTICE\_DIM.

Some of the columns were joined and made into single column:

* PRACTICENAME was actually the address where the locum was practising so PRACTICENAME, ADDRESSLINE1 and ADDRESSLINE2 were joined and named ADDRESS.

Command to create this dimension is:

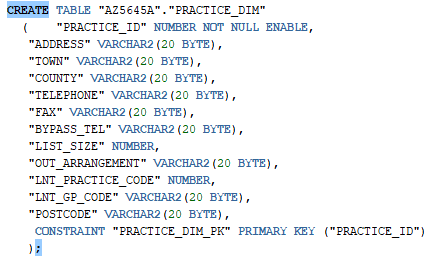


Figure 17

## DIMENSION 5: TIME\_DIM table

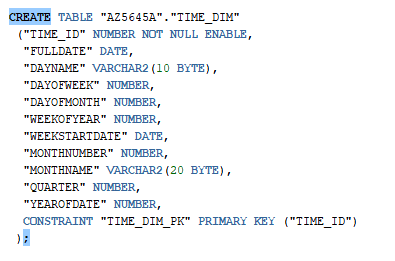
Time dimension table is undoubtedly the most important table as many required queries needed data in relative to the time so I decided to build an extensive time dimension according to our needs and data given. Command to create this dimension is

Figure 18

## FACT 1: SESSIONS\_FACT table

This is one and only fact table in our star schema. This table is basically created by combining all the necessary columns of SESSION table and all the keys from 5 dimension tables for referencing purposes. 5 foreign keys in this fact table are – REQUEST\_ID, TYPE\_COVER\_ID, LOCUM\_ID, PRACTICE\_ID and TIME\_ID.

SQL command to create the fact table is:

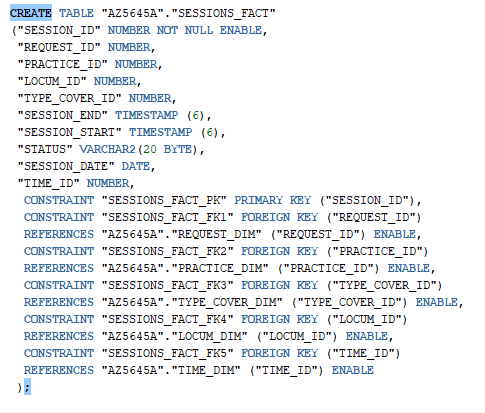


Figure 19

# Data cleaning plan and practices

Before loading data into the newly created star schema, it needed to be cleansed so that it can yield authentic reports. After importing all the required tables in SQL developer from Microsoft Access, our staging process was complete. Now important tables which needed data cleansing were – LOCUMDETAILS, PRACTICEDETAILS and SESSION. Cleansing and making data consistent in the mentioned tables was very important as data from these tables was supposed to be loaded into the data warehouse.

Data needed to be consistent and a lot of issues required to be dealt with. Some of the issues were:

* ***Missing data:*** Either replaced manually or the column dropped if too many fields where empty.
* ***Null values:*** If this was in a required field, then a default string was used.
* ***Redundant columns:*** For example, there were too many columns representing the address, so I merged them into one column showing all the necessary information.
* ***Tables referring to other tables for simple information:*** An example of this is the TYPE\_OF\_COVER table which instead of a description such as “Surgeries” or “On call” , shows numerical values such as 1 and 2 which were used then to reference to another table. This made the information incomprehensible or difficult to understand and an approach was taken to replace these key values with the actual data.
* ***Duplicated data***: Any columns that had duplicate information where deleted or merged.
* ***Formatting date and time***: After the data export from MS Access to Oracle SQL developer, some time and date records where formatted incorrectly so this data was all formatted.

## Cleaning table 1: LOCUMDETAILS

In the cleaning process, only those column sets which were required to load the data warehouse were cleaned and made consistent.

* First Name and Last name columns were joined in the First Name column.
* Manually from SQL developer removed unnecessary spaces in “First Name” column.
* Address Line 1 and Address Line 2 were joined in the Address Line 1 column.
* Filled all the null values in “town” column with respect to data in “Address line 2” column.
* All the remaining null values in “Town” were filled with ‘Greenwich’.
* Some values in “Town” were adjusted manually.
* Filled all the null values in “county” column with respect to data in “town” column.
* All remaining null values in “county” column were filled with “London”.
* All values of 0 and null in “gender” column were replaced with ‘not known’.
* All the “postcode” values were formatted correctly on SQL developer.
* All null “postcode” values were replaced with ‘unknown’.
* All “Tel#6ome” values were set to ‘1111 111 1111’ which were null.
* “Date application was sent” column was filled with date of ’01-JAN-01’ where there were null values.
* “Date of birth” column was converted from Varchar to Date format.
* All null Date of birth values were filled with ’01-JAN-01’.
* “Nationality” column was also filled with ‘Not known’ in place of null values.
* Columns that need to be translated: “**Gender”**, **“Current Status” and “Nationality”.**

Data Cleansing screenshot for this table is on next page:

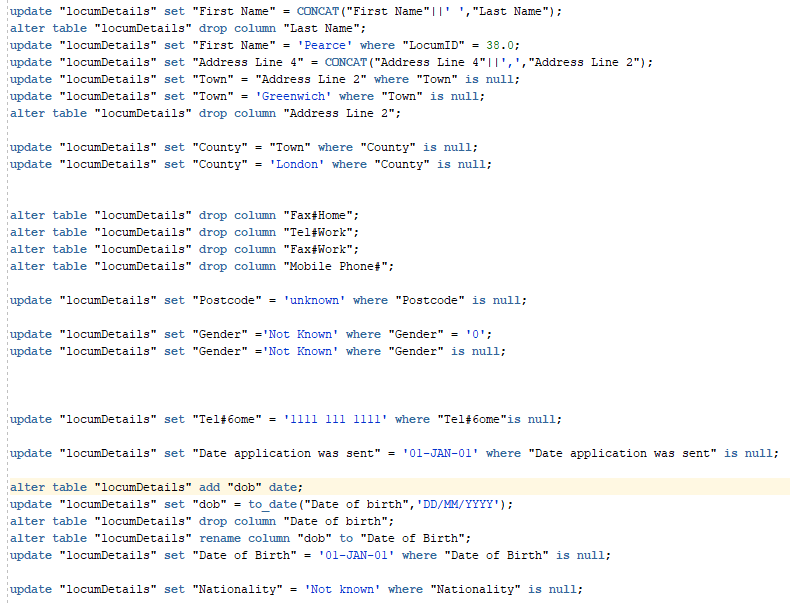


Figure 20

**NOTE**: This code is also attached with the final submission.

## Cleaning table 2: PRACTICEDETAILS

Only the columns which were required to load PRACTICE\_DIM were cleaned from PRACTICEDETAILS.

* “practice name”, “address line 1” and “address line 2” were joined because all these 3 fields were representing address of practitioner.
* Some of these columns needed a manual inspection and editing for example addition removal of commas and spaces.
* The “Address Line 1” column name was changed to “Adress\_Line\_1” because the previous name was giving error of invalid identifier whenever using it with WHERE clause.
* Rows which had null entry in most of there fields were removed. i.e. row with PracticeID of 57.
* All the rows with empty “Town” field were filled accordingly with relevant information.
* All the rows with empty “County” field were filled accordingly with relevant information.
* Empty “Fax#” fields were filled with “1111 111 1111”.
* Empty “Bypass#” fields were filled with “1111 111 1111”.
* Empty “Approximate List Size” fields were filled with “Not known”.
* Empty “Out of Hours Arrangement” fields were filled with “Not known”.
* Empty “LNT practice code” fields were filled with “Not known”.
* Empty “LNT GP code” fields were filled with “Not known”.

The code for above explained changings:



Figure 21

**NOTE**: This code is also attached with the final submission.

## Cleaning table 3: SESSION

This is the most valuable table in the whole database, so it requires more care than the others. Data from this table is supposed to be made ready so that it can be transferred into our SESSION\_FACT table.

* This table contains a lot of foreign keys, and each row should have every foreign key in order to get referenced in the future. All the rows in which one or more than one foreign keys were null or ‘0’ were either deleted or the foreign value was replaced with a decided value.
* “LocumID” was null or ‘0’ in 219 rows so I could not delete this many rows of data. In order to preserve the referential integrity, 0 and null value “LocumID” rows were replaced with 11 (decided value).

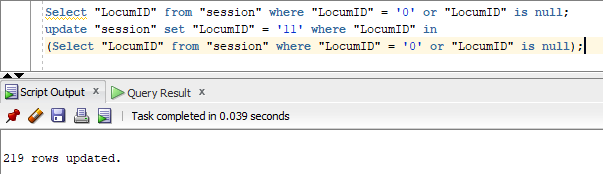


Figure 22

* All the rows which has value of 0 in “TYPE” column were replaced with 5 (decided value) to maintain the consistency of database. This “type” column is going to be referenced with “Type\_cover\_ID” in TYPE\_COVER\_DIM.

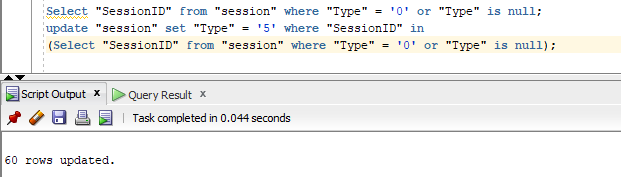


Figure 23

NOTE: The tables Template\_type\_of\_cover and LocumRequest did not need any cleaning of the columns which were required for loading the star schema.

# Database Mapping

This section is basically to help load the data into newly created star schema more easily and efficiently. It marks down the mapping of columns from source to destination.

1. SESSIONS\_FACT ------- session

-SESSION\_ID = SessionID

-REQUEST\_ID = RequestID

-PRACTICE\_ID = REQUEST\_DIM.PRACTICE\_ID (After loading REQUEST\_DIM)

-LOCUM\_ID = LocumID

-TYPE\_COVER\_ID = Type

-SESSION\_END = SessionStart

-SESSION\_START = SessionEnd

-STATUS = Status

-SESSION\_DATE = SessionDate

-TIME\_ID = TIME\_DIM.TIME\_ID (After loading TIME\_DIM)

1. LOCUM\_DIM ------- LocumDetails

-LOCUM\_ID = locumID

-TITLE = Title

-NAME = First Name

-ADDRESS = Address Line 4

-TOWN = Town

-COUNTY = County

-POSTCODE = Postcode

-TELEPHONE = Tel#6ome

-APPLICATION\_DATE = Date application was sent

-GENDER = Gender

-CURREST\_STATUS = Current Status

-DOB = Date of Birth

-NATIONALITY = Nationality

1. REQUEST\_DIM ------- LocumRequest

-REQUEST\_ID = LocumRequestID

-PRACTICE\_ID = PracticeID

-REQUEST\_DATE = Request date

1. TYPE\_COVER\_DIM ------- Template\_Type of cover

-TYPE\_COVER\_ID = TypeofCoverID

-DESCRIPTION = CoverDescription

1. PRACTICE\_DIM ------- practiceDetails

-PRACTICE\_ID = PracticeID

-ADDRESS = Address\_Line\_1

-TOWN = Town

-COUNTY = County

-POSTCODE = Postcide

-TELEPHONE = Telephone#

-FAX = Fax#

-BYPASS\_TEL = Bypass#

-LIST\_SIZE = Approximate List Size

-OUT\_ARRANGEMENT = Out of Hours Arrangement

-LNT\_PRACTICE\_CODE = LNT practice code

-LNT\_GP\_CODE = LNT GP code

1. TIME\_DIM ------- session.Session\_Date and REQUEST\_DIM.Request\_date

# Data Loading

After completing the data cleansing process, the cleaned data was loaded into the newly implemented star schema.

In order to have a successful load and to meet the project’s requirement specification, cursor was used to load the data. Cursors are pointers to areas where information about data manipulation language (DML) statements are processed (Feuerstein, 2013). Cursors are an efficient and easy to use tool used to query and change the contents of tables in a database.

In the following sections, cursors are used to load cleansed data from the staging area in SQL Developer into the LOCUM\_DIM, PRACTICE\_DIM, REQUEST\_DIM, TYPE\_COVER\_DIM and TIME\_DIM dimension tables as well as the fact table SESSIONS\_FACT.

**Note:** All the codes in the following sections are attached in the zipped file in the final submission.

## Step 1: Using cursor to load LOCUM\_DIM

After declaring and running the cursor c\_LocumDetails, the following columns, "Address Line 1","Town","County","Postcode","Telephone#","Fax#","Bypass#","Approximate List Size","Out of Hours Arrangement","LNT practice code","LNT GP code", were loaded from the LocumDetails table into corresponding columns in the LOOCUM\_DIM table.

Since the data type of the CURRENT\_STATUS column which was mapped to the LocumStatus column in the LocumDetails table was binary\_double, and did not agree with that of the fact table, an additional column, CURRENT\_STATUS2 of data type varchar2, was created after the data loading was completed. It was then updated with values from the CURRENT\_STATUS table to ensuring the data type matches with that of the fact table.

The GENDER column of the LOCUM\_DIM table was updated by mapping it with data found in the Template\_Gender table in the MS Access database provided for the project where ‘1’ = ‘female’, and ‘2’ = ‘male’.

Similarly, the CURRENT\_STATUS2 and NATIONALITY columns were updated by mapping them to data found in the Template\_Locum Current Status and Template\_Nationality, respectively.

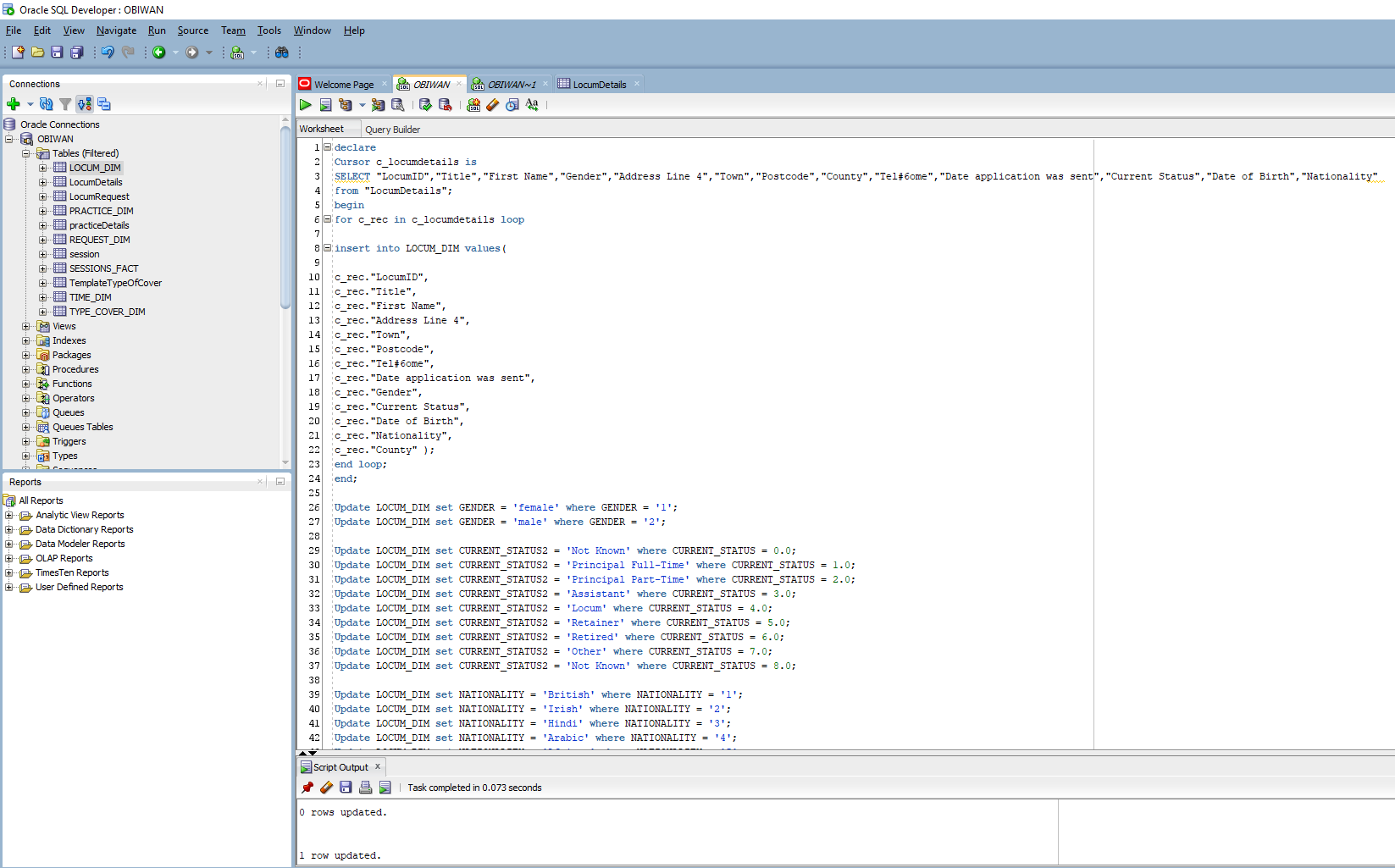


Figure 24: Screenshot of cursor loading data into LOCUM\_DIM table

## Step 2: Using cursor to load PRACTICE\_DIM

Similar to step 1 above, the cleansed data from the columns, "PracticeID","Address Line 1","Town","County","Postcode","Telephone#","Fax#","Bypass#","Approximate List Size","Out of Hours Arrangement","LNT practice code","LNT GP code", from the praticeDetails table were loaded using the cursor c\_practiceDetails into the PRACTICE\_DIM table. The was no updates in this case as the data type of the columns matched with that of the fact table.

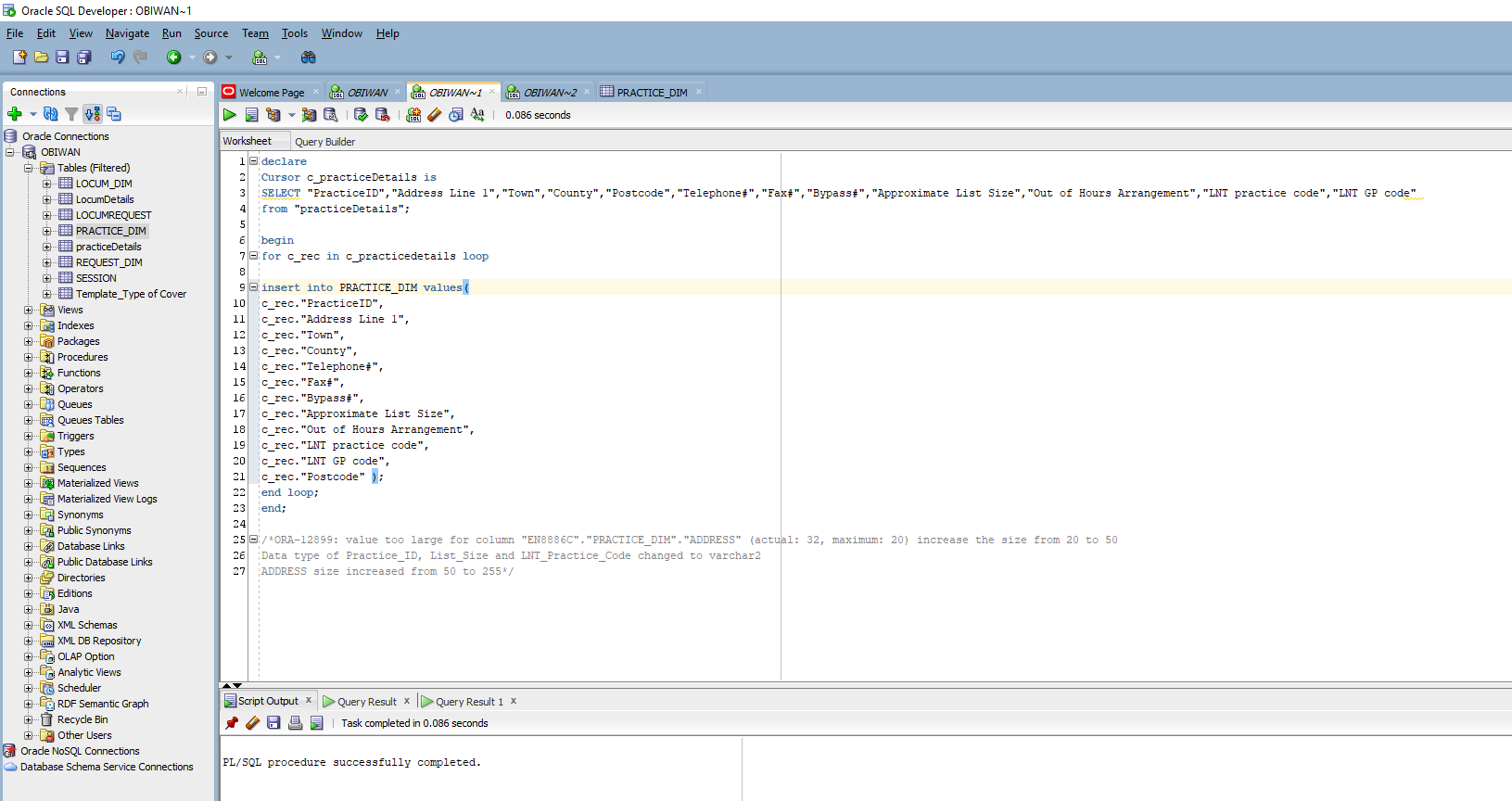


Figure 25:Screenshot showing data loaded into the PRCTICE\_DIM table

## Step 3: Using cursor to load REQUEST\_DIM

Like the previous steps, the columns, "LocumRequestID","PracticeID","Request date", were loaded from the LocumRequest table to the REQUEST\_DIM table using the c\_locumRequest cursor.

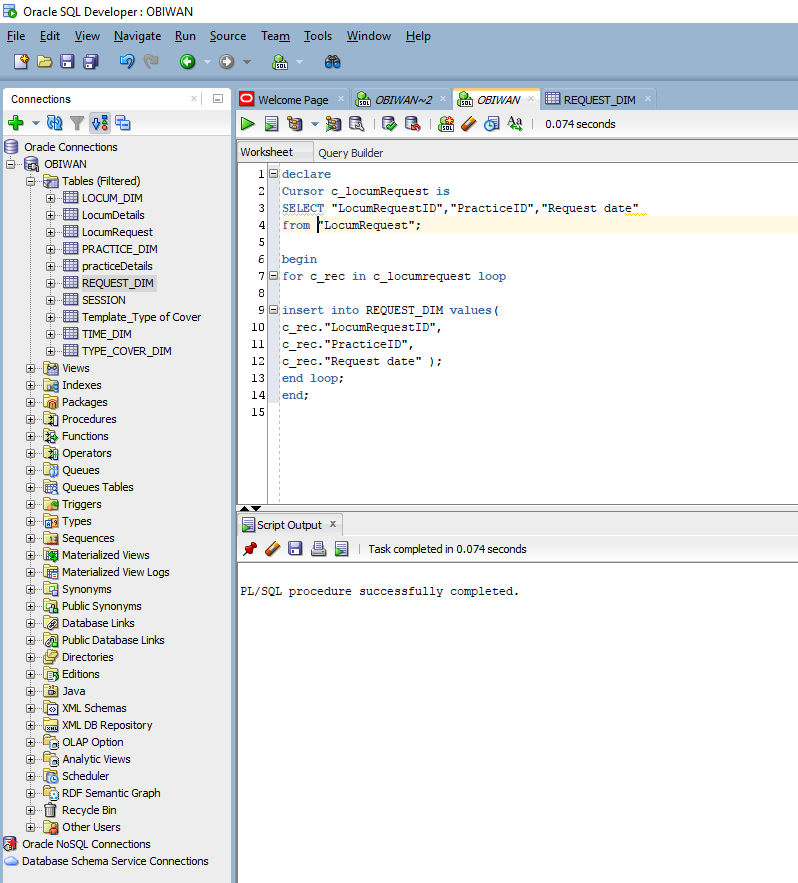


Figure 26: Loading data into the REQUEST\_DIM table from the LocumRequest table

## Step 4: Using cursor to load TYPE\_COVER\_DIM

Like the previous steps, the columns, "TypeofCoverID","CoverDescription", were loaded from the TemplateTypeOfCover table to the TYPE\_COVER\_DIM table using the cursor c\_TemplateTypeOfCover.

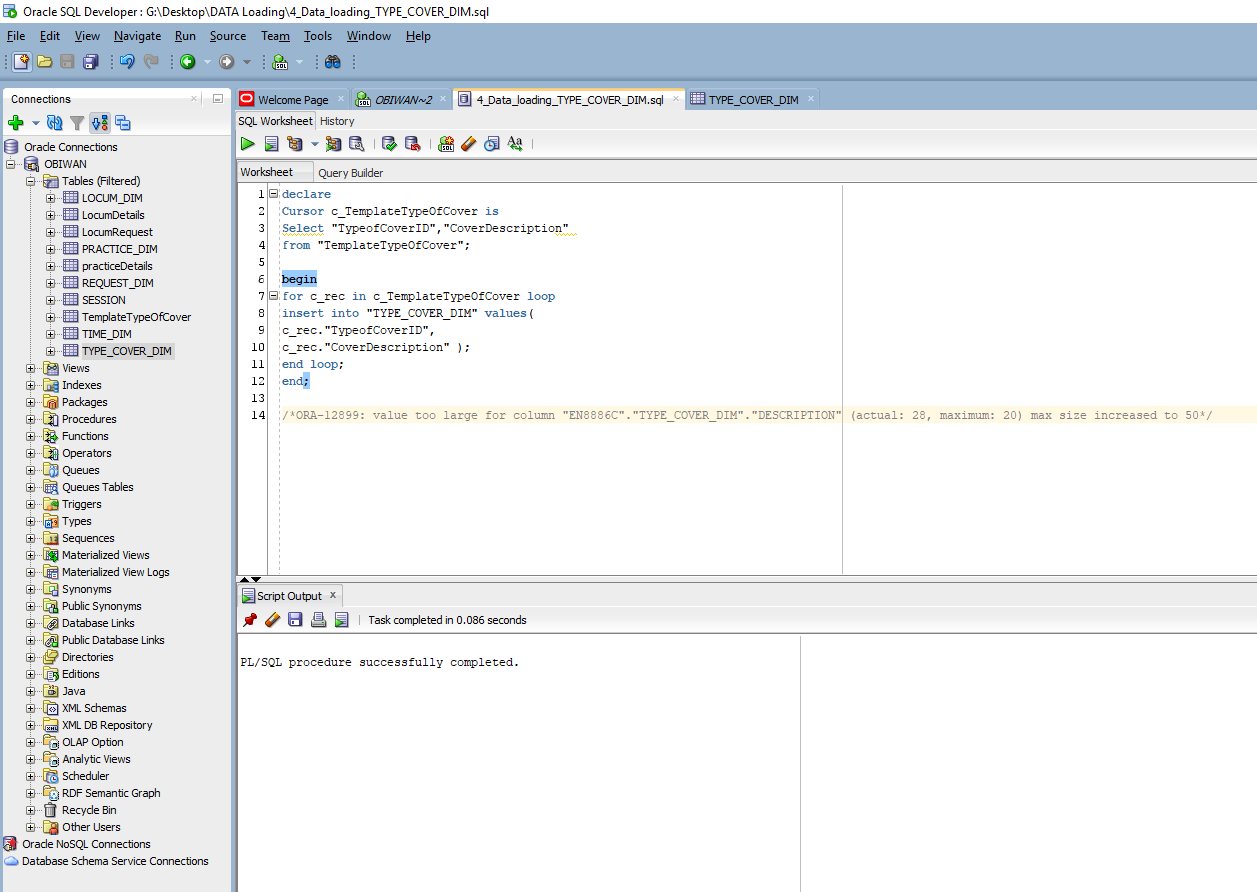


Figure 27: Loading data into the TYPE\_COVER\_DIM table from the TemplateTypeOfCover table

## Step 5: Loading TIME\_DIM

For the TIME\_DIM loading, the SessionDate and REQUEST\_DATE columns were needed from the session and REQUEST\_DIM table, respectively.

A sequence is used to populate the TIME\_ID column while the data loaded from the SessionDate and REQUEST\_DATE columns is split into the columns FULLDATE, DAYNAME, DAYOFWEEK, DAYOFMONTH, WEEKOFYEAR, WEEKSTARTDATE, MONTHNUMBER, MONTHNAME, QUARTER, YEAROFDATE.

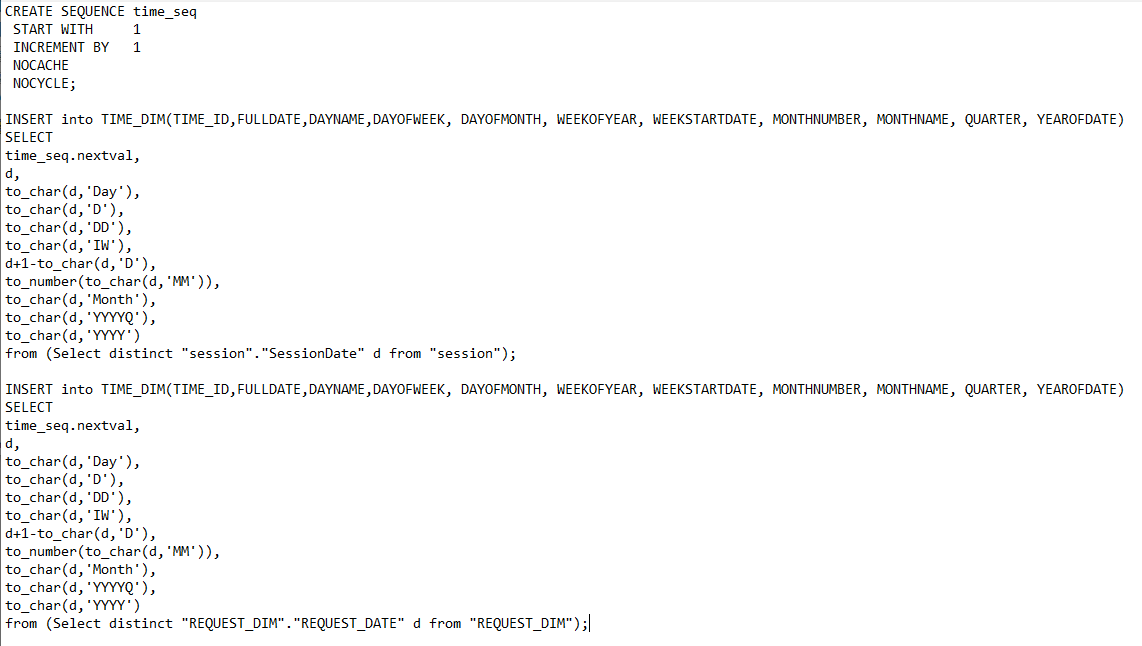


Figure 28: Loading TIME\_DIM table

## Step 6: Using cursor to load SESSION\_FACT

Finally, we arrived at the stage where we had to load the SESSIONS\_FACT table of our Star schema. Most of the columns were taken from sessions table- SessionID, RequestID, LocumID, Type (ID), SessionStart, SessionEnd, Status, SessionDate. Two of the columns PRACRTICE\_ID and TIME\_ID were taken from REQUEST\_DIM and TIME\_DIM, respectively. Below you will see the code used for loading the SESSIONS\_FACT table:

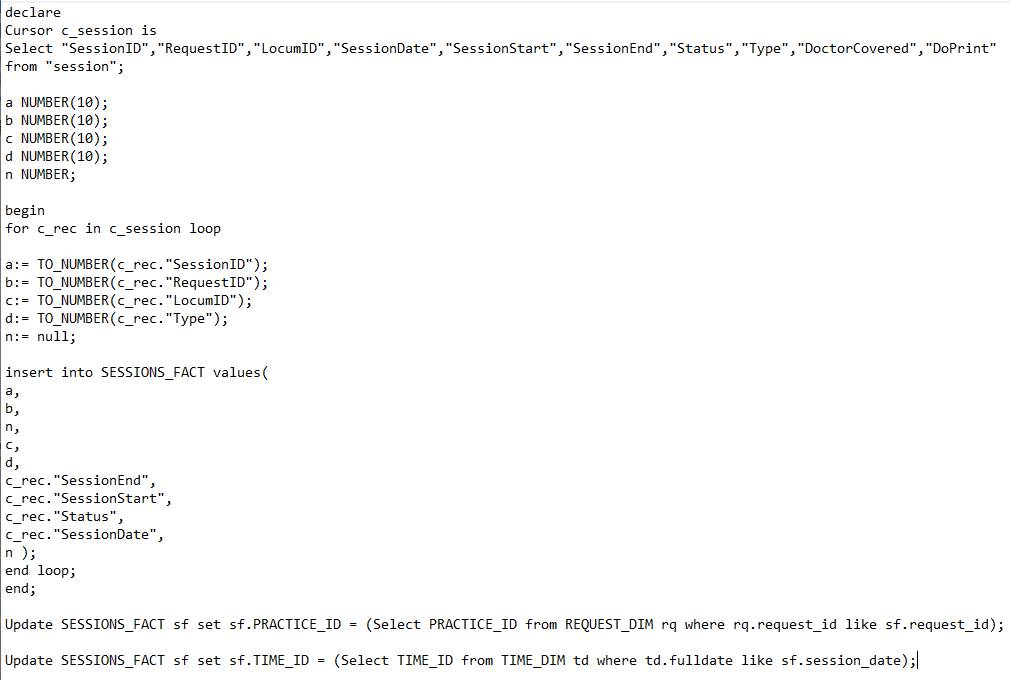


Figure 29: SESSIONS\_FACT loading

# Information Queries

Below you will see all 6 SQL queries to get the reports which were required by NHS. Each of the queries are written with exceptional attention to detail in order to display the most meaningful report possible.

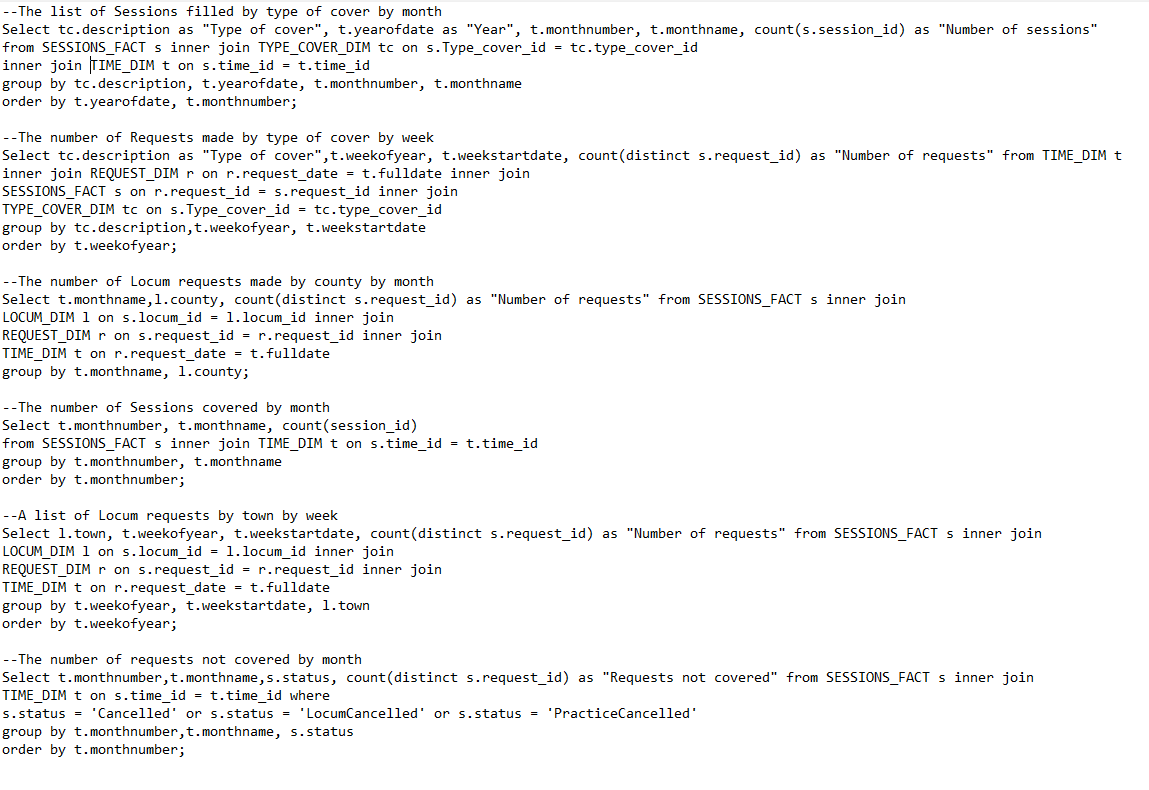


Figure 30 - Requirement Queries

NOTE: This code is also attached with final submission in the information queries folder.

# Summary

All the requirements needed by NHS were fulfilled in this project. Only the required tables were considered in order to load the designed star schema. The required table columns were cleansed in the staging area properly before loading into Star schema. All 6 information queries represent the reports required by NHS as specified in the course work specification. Now these queries can be run anytime, or a personalised view might the created for each of the queries for efficient report generation.