|  |  |
| --- | --- |
| **C:\Users\sram\OneDrive - Unitec NZ\ShiuRam\Unitec\Unitec Logos\UnitecHorizontalLogo.jpg** | ISCG6425 Data Warehousing  **Assignment 2**  **Semester 2, 2022** |
| School of Computing, Electrical and Applied Technology | **Due Date:** 9 November 2022  **Deadline Time:** 8:30 pm |
|  | **Total Marks:** 100  **Course Weighting:** 30% |

This is **individual** assignment.

**Purpose of Assignment**

* Develop SQL script using appropriate techniques to perform ETL process
* Develop SQL script to perform Data Quality checking and logging
* Implement a complete data warehouse solution using Integration tool

**Specifications and Instructions**

In year 2012, Models Sales Company had MySQL database management systems (OLTP database, **salemodels**) to store their transactional data related to their customers, products, and sales transactions. Later they bought company that run a same business in Pacific region. That company has MSSQL database **saleAU\_NZ.**

Now, the Sales department manager wants to integrate sales data stored in the two databases into a single database to help them have a central access to all data and perform data analysis. Therefore, they have decided to implement a data warehouse for this purpose. Assume that you have been employed as a database analyst for the company and you are assigned to carry out the following tasks to design and populate a data warehouse for the company.

Please note that saleAU\_NZ was implemented on MSSQL management studio and salemodels was implemented using MySQL

Fortunately, the same schema is designed for both salemodels and classicmodels, as shown in Figure 1y, only slightly different is in the tables name. Please check.

The **Profitability** (P) of a product can be defined as follows:

P = [Order Details].PriceEach– Products.buyPrice

If P > 0, the product makes profit; If P < 0, the product makes loss.

Important Totals for Fact table:

[TotalPrice]=(od.QuantityOrdered\* od.PriceEach)

[TotalProfit]= (od.QuantityOrdered \* (od.PriceEach - p.buyPrice))

[TotalPossibleProfit] =(od.QuantityOrdered \* (p.MSRP - p.buyPrice))

(p- Product table, o- OrderTable, od -Order Detail Table)

The manufacturer's suggested retail price (MSRP) is the price that a product's manufacturer recommends it be sold for at point of sale.

Please note that same product sold price is different in different order details, as different discounts for the same product was run in the different time frames.

Please be mindful that total sales of a product is equal to the total Quantity from all of the orders. The total number of products sold/bought refers to how many different ProductIDs are sold/bought.

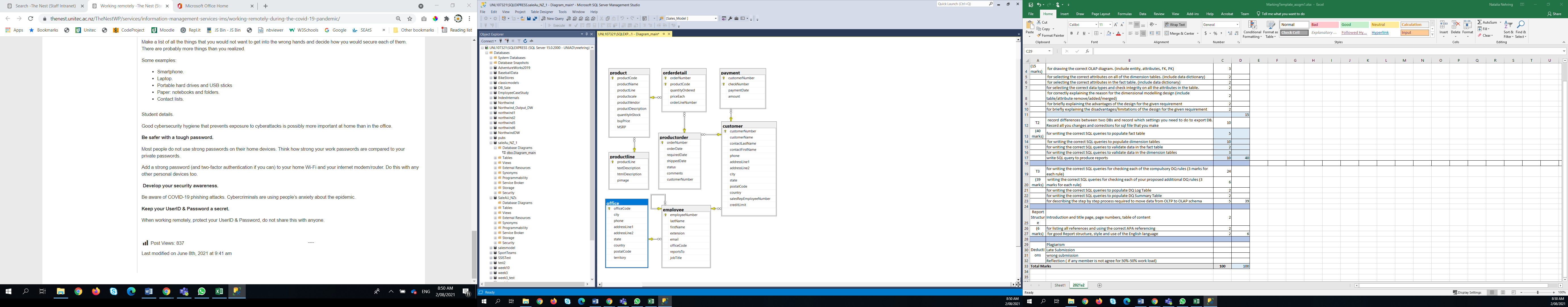


Figure 1: OLTP Schema for saleAu\_NZ (MSSQL)

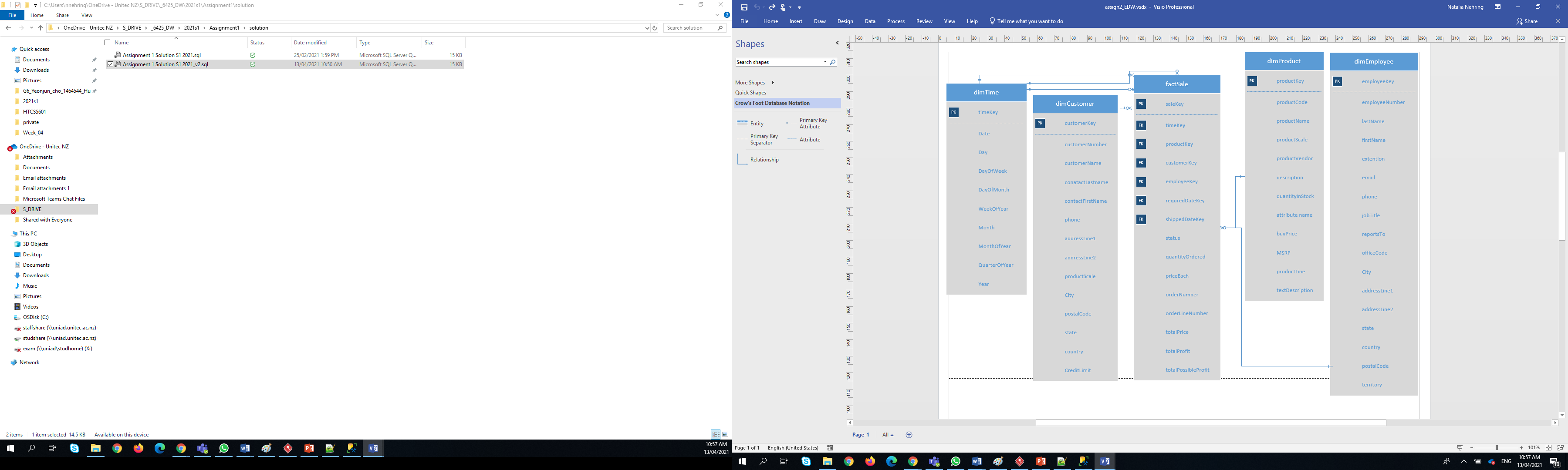


Figure 2: Figure 2. OLAP Database Schema

**Tasks T1 OLAP and ETL process**

1. Create your OLAP database based on star schema, you can use your Assignment 1 solution.
2. Use MSSQL for both DB, you can use your databases from Assignment 1.
3. You are required to create a new Integration Services project (Visual Studio). Using tool - Integration Services project - extract, transform and load data from both databases to your OLAP database.
4. Record your Integration Services project development as progress log

All data in the OLAP database must meet the following criteria.

1. All data from *all* tables of the source databases must correctly appear in your OLAP database.
2. All PKs and FKs must be valid.

## Describe the ETL and evaluation process that you have done to ensure that your proposed design has met all of the requirements above.

## You need the total number of data in OLAP vs. OLTP, etc.

**T2 Data Quality**

1. Generate Data Quality Rules to ensure that all of the data imported to the Data Warehouse are clean, i.e. they are all accurate, valid, consistent, complete and uniform. It must also include the Data Quality (DQ) Rules mentioned in Table 1. You are also required to add two additional DQ rules of your choice.
2. Check each of the data from *saleAU\_NZ* and *salemodels)* that needs to be imported to the OLAP database against the Data Quality Rules. Log all records that violate the Data Quality Rules that you have created on Task 2 to a *DQLog* table and summarize it in a *DQ Summary report*.
3. Import the clean data from *saleAU\_NZ* and *salemodels)* to the OLAP database.
4. Check that the data that you have imported are all correct.

**Table 1. Data Quality Rules**

|  |  |  |
| --- | --- | --- |
| **Rule No.** | **Description** | **Action (If-Then)** |
| 1 | **buyPrice** checking in Products | **Reject** if **buyPrice** is 0or Null  **Fix** if **buyPrice** is negative, need to convert to positive number |
| 2 | **priceEach** checking in orderDetails | **Reject** if **priceEach** is 0or Null  **Fix** if **priceEach** is negative, need to convert to positive number |
| 3 | **QuantityOrdered** checking in Order Details | **Reject** if **Quantity Ordered** is Null or 0  **Fix** if **Quantity Ordered** is negative, need to convert to positive number |
| 4 | **MSRP** checking in Product | **Reject** if **MSRP** < **buyPrice** on a product |
| 5 | **Country** checking in *Customers* and *Offices* | **Fix** the *Country*format *regardless* capital or non-capital cases, e.g.  US, United States, UNITED STATES 🡪 USA (in saleAU\_NZ)  Australia 🡪 AU  New Zealand 🡪 NZ |
| 6 | **ProductCode** checking in *Order Details* | **Reject** if*ProductCode* doesn’t exist or is *null* |
| 7 | **CustomerNumber,** addressLine1, addressLine2 andCity checking in *Customers* | **Reject** if(*CustomerNumber* doesn’t exist or is *null*) **and** (if any of: addressLine1, addressLine2 andCity are *null*) |
| 8 | **requiredDate, shippedDate** checkingin *Orders,* | if*orderDate* greater than **requiredDate,**  **Fix if requiredDate null replace** *orderDate+7*  if*orderDate* greater than **shippedDate**  **Fix if shippedDate null replace** *orderDate+2, if status is ='Shipped'* |

**Presentation**

You will be required to present your solution design and rationale, key technique/technology used, and conduct a live demo of your solution to the class. Durations are limited to 10 minutes for the presentation and the demo. You will be required to answer 2 - 4 random questions regarding the work. This is to verify that you have understood the assignment and undertaken the work by yourself.

**Assignment Submission**

* Your submission folder must include next files:
  1. An **SQL file** containing SQL codes for DB salesmodel after clean up and upload to MSSQL – 1\_*salesmodel\_MSSQL.sql*
  2. An **SQL file** “2\_*Create\_ STAR\_schema.sql* “ containing SQL codes to populate DW solution.
  3. Integration Project **folder - named** 3\_*IntegrationSP\_Assignment2* **Task T1** and evaluation on the validity of your design and ETL process **Task T2.**
  4. An **SQL file** containing SQL codes for DQlog file, Data Logging and summary tables. 4\_DQlog.sql
  5. Report (Word document 5\_report.docx):
     1. Tasks T1 Explain your steps to create your Integration project
     2. Task T3 Data Quality -explain reasons for each rule. Present you log file.
* Put all files in one folder. Please rename this folder – “Student1\_Name1\_ID1\_ Student2\_Name2\_ID2”
* Zip this folder- using a ZIP program only.
* Only one member of your team is required to submit the ZIP folder on Moodle
* If you are not agreed for 50%-50% contribution to this assignment from each group member, you will need upload reflection report (sample available on Moodle).
* You must acknowledge and declare on the coversheet of assignment (download from Moodle) that your submitted work has been done by yourself and shows no attempt of plagiarism. If any cases of plagiarism are suspected, you are required to declare the ownership and authenticity of your work to the marker. **Maximum mark deduction** may be applied to your work.
* For the purposes of academic integrity, students who haven’t demonstrated progress work in the class time (and/or no check point submission) can be asked to demo/test their working code and explain logic to the lecturer individually after assignment submission.

**Marking Schedule**

|  |  |  |
| --- | --- | --- |
| **Task** | **Marking Criteria** | **Marks** |
| **T1** | Star schema: |  |
| **(57 marks)** | Create your OLAP database based on star schema. | 2 |
|  | Use MSSQL for both DB, you can use your databases from assignment 1. | 2 |
|  | Populating dimensional tables. From both DBs | 5 |
|  | Populating fact table. From both DBs | 4 |
|  | Use calculate – derived Column | 5 |
|  | Use Merge for join data from 2 tables | 4 |
|  | Development log -should have as minimum 5 different version of development explained (screen shorts, description, project compiled, version saved) | 35 |
|  |  |  |
| **T2** | for writing the correct SQL queries for checking each of the compulsory DQ rules (3 marks for each rule) | 24 |
| **(31 marks)** | writing the correct SQL queries for checking each of your proposed additional DQ rules (3 marks for each rule) | 6 |
|  | for writing the correct SQL queries to populate DQ Log Table | 1 |
|  |  |  |
| **Report Structure** | Title page, introduction, page with statement, that it is own work. | 1 |
|  | for good Report structure, page numbers, table of content, style and use of the English language | 1 |
|  |  |  |
| **Presentation**  **(10 Marks)** |  | 10 |
|  |  |  |
| Deductions | Plagiarism (up to 100 deduction) |  |
| Late Submission (explained next) |  |
| Not follower submission instruction, (up to -10) |  |
| **Total Marks** | | **100** |

**Presentation Marking**

Name of Learner:

Name of Observer:

Date of Presentation:

|  |
| --- |
| Indicate if the Learner has met the criteria during the process of achieving the objective. Use the space to add comment for feedback to the Learner and for moderation purposes. |

|  |  |  |
| --- | --- | --- |
| Observation | Comment | Mark |
| Appropriate body language (inc movement and gestures) |  | 1 |
| Appropriate usage of voice, pronunciation and volume |  | 1 |
| Language used is suited to the audience (fits purpose, audience, and context?) |  | 1 |
| Subject knowledge |  | 2 |
| The subject matter was organised logically |  | 2 |
| Effective use of visual aids |  | 1 |
| Answering questions(2 - 4) |  | 2 |
| **TOTAL** |  | **10** |
| Extra Notes: | | |

Observer’s Signature:

**Late Submission of Assignments:**

Assignments submitted after the due date and time without having received an extension through Special Assessment Circumstances (SAC) will be penalised according to the following:

* 10% of marks deducted if submitted within 24hrs of the deadline,
* 20% of marks deducted if submitted after 24hrs and up to 48hrs of the deadline,
* 30% of marks deducted if submitted after 48hrs and up to 72hrs of the deadline,
* No grade will be given for an assignment that is submitted later than 72hrs after the deadline.

Affected Performance Consideration

A student, who due to circumstances beyond his or her control, misses a test, final exam or an assignment deadline or considers his or her performance in a test, final exam or an assignment to have been adversely affected, should complete the APC application found at <https://www.unitec.ac.nz/current-students/study-support/affected-performance-consideration>

Within any semester, a student may have only one APC per course.

When requesting an APC for an assignment, the APC must be submitted (along with work completed to-date) within the time frame of the extension requested; i.e. if the Doctor’s certificate is for one (1) day, then the APC and work completed must be submitted within one (1) day.

**Assistance to other Students:**

Students themselves can be an excellent resource to assist the learning of fellow students, but there are issues that arise in assessments that relate to the type and amount of assistance given by students to other students. It is important to recognise what types of assistance are beneficial to another’s learning and also what types of assistance are unacceptable in an assessment.

**Beneficial Assistance:**

* Study Groups
* Discussion
* Sharing Reading Material
* Reading the available online and library resources

**Unacceptable Assistance:**

* Working together on one copy of the assessment and submitting it as own work
* Giving another student your work
* Copying someone else’s work, this includes work done by someone not on the course
* Changing or correcting another student’s work
* Copying from books, the Internet etc. and submitting it as own work; anything taken directly from another source must be acknowledged correctly; show the source alongside the quotation