Contents

Word Count: 2205

1. EXECUTIVE SUMMARY

1.1 Content summary

This report mainly aims to state and explain the design issues about ETL (Extraction, Transformation and Loading) process based on the given data. In addition, this report discusses the redesign of data warehouse to accommodate data properly. The last part talks about additional data dictionary used in the ETL process.

1.2 Overview about ETL process

Before the ETL process, source data could come from various sources, such as flat files, relational database and external files. Then it will be extracted to form a new table. The next step is to transform data so that it could keep uniform. The last step in staging area is that data waits to be loaded completely to the target data warehouse.

From the horizontal view, there are source systems, staging area and data storage area. They will be visited in sequence as the development of ETL process. Moreover, a reporting layer outside the data warehouse server could support the process of business intelligence, such as analysis for business profit and proposing business reporting.

The following figure signifies the whole process of the establishment and employment of one data warehouse.

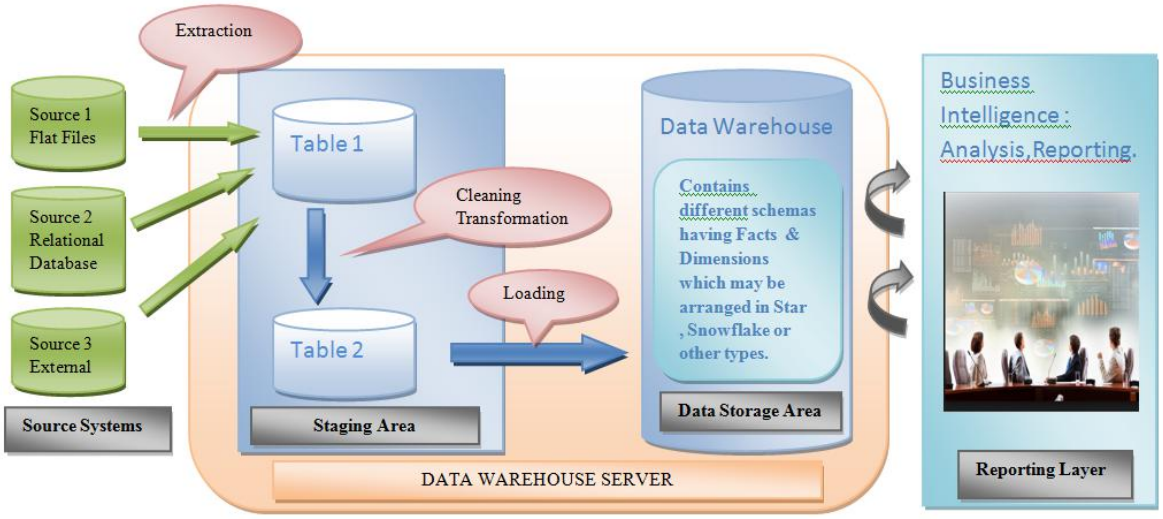


Figure 1 overview about the process of data warehouse

Source: (testingpool, 2015)

1.3 Problems addressed in designing ETL process

The first issue to be overcome is poor and inconsistent data quality. As the selection of different software and the existence of individuals' preference, data form and data type from various sources could be inconsistent. Hence, controlling data quality is a basic issue in

ETL process. It could simplify the process and reduce unnecessary work load in subsequent sections.

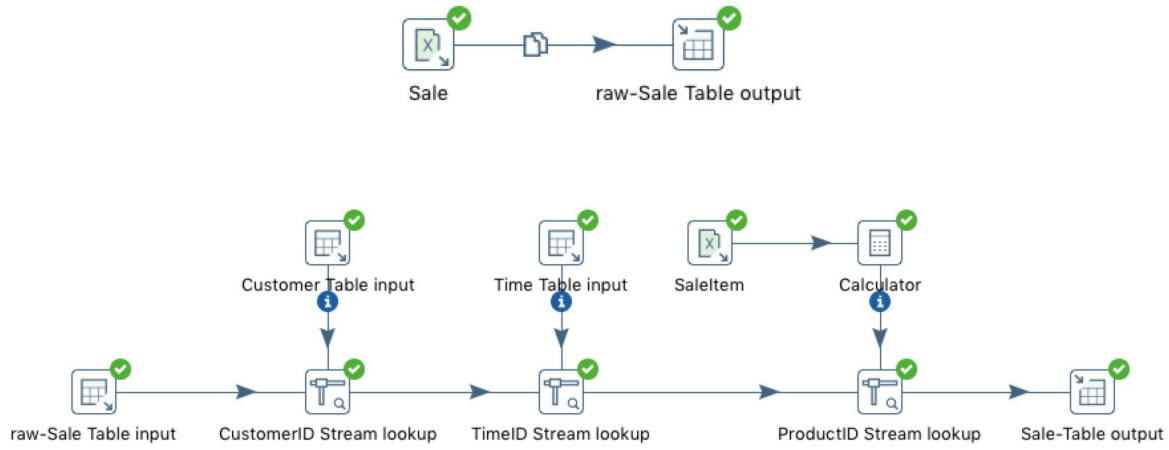
The second issue is the removal of null-value data. The first step to solve this problem is to recognise all null-value data and then delete them. The reason is to save processing time and storage room.

And the last one needs to be addressed is the slowly changing elements. For example, cost of each product is a slowly changing element in Product Table, while postcode is recognised as a slowly changing element in Customer Table. Kimball (2008) presents several methods to handle slowly changing dimensions. This ETL process chooses the method of Type 2. In detail, two new incremental columns are added respectively so that they could uniquely identify each product under each cost and postcode for each customer during certain time periods.

2. Design of the ETL Process

ETL process is the efficient method to extract the source data and transform it into the predesigned data warehouse. As designed in Assignment 1, the data warehouse has four dimensions and one fact table. Each of them has their own transformation. This part will introduce these five processes in detail.

2.1 Sale Transformation



Sale table is the fact table of the data warehouse. Therefore, it has four outputs which refer to each four dimensions.

First, we transformed the source data ‘Sale.xlsx’ into a MySQL table called raw-Sale table. This quick pre-process is in order to reduce the time of execution. Since this table has more

than 220,000 pieces of data, using MySQL table as an input is much more efficient rather than the Excel file.

What's more, as the fact table of the star schema, this Sales fact table needs to store all other dimensions' primary key as its foreign key. That is to say, this transformation is a cooperation of all tables. In the above figure, in order to reduce the executing time and avoid doing duplicated work, we take the Customer table and Time table, which are the output of the Customer dimension and Time dimension, directly as the input.

And then, as shows above, from left to right, we merge the attributes step by step. Using the Stream lookup step, we first add the CustomerID to the Sale table, the the TimeID, and then the ProductID.

As the data warehouse designed, the stakeholders need to access to the raevenue so that they can understand the margin easily. Therefore, the Calculator step is used for the revenue. Lastly, we output the Sale table as a MySQL table for the future analysis.

This transformation is executed successfully. And the result of this process is as follows.

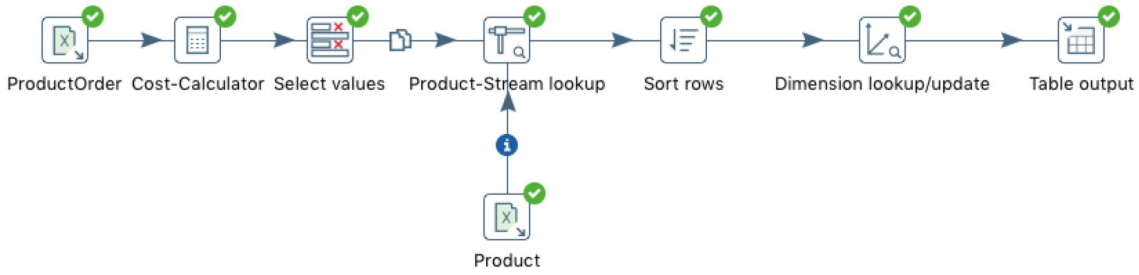
SaleID	ProductID	CustomerID	TimeID	StoreID	Revenue	Quantity	Sale_Price
S1	100867	8	1	1	49.08	1	49.08
S2	100340	27	1	3	131.55	1	131.55
S3	100182	NULL	1	2	149.75	1	149.75
S4	101117	41	1	3	98.10000…	5	19.62
S5	100717	NULL	1	2	92.08	1	92.08
S6	100566	NULL	1	3	215.26	2	107.63
S7	100492	1321	1	1	232	2	116
S8	100386	NULL	1	2	129.16	1	129.16
S9	101505	NULL	1	2	409.32	2	204.66
S10	101334	118	1	1	926.67	3	308.89
S11	100794	NULL	1	2	65.8	1	65.8
S12	101269	NULL	1	2	395.6	1	395.6
S13	101043	NULL	1	2	50.2	2	25.1
S14	100618	227	1	3	501.55	5	100.31
S15	100322	230	1	1	131.55	1	131.55
S16	100909	NULL	1	2	86.94	2	43.47
S17	100504	NULL	1	3	232	2	116
S18	101643	NULL	1	2	356.9	2	178.45
S19	100560	NULL	1	2	215.26	2	107.63
S20	100637	292	1	1	95.67	1	95.67
S21	101607	NULL	1	3	178.45	1	178.45
S22	101309	337	1	1	318.12	1	318.12
S23	100217	NULL	1	3	141.12	1	141.12
S24	100954	NULL	1	3	36.8	1	36.8
S25	100659	407	1	3	191.34	2	95.67

Sale 1

Sale table

As it shows above, for each SaleID, there are ProductID, CustomerID, TimeID, and StoreID as the foreign key linked to other dimensions. After executed, we found that some fields in CustomerID are NULL, which means that some CustCode of the Sale.xlsx do not exist in the Customer Table. This refers to a data quality issue. And it should be solved in the future.

2.2 Product Transformation



Product table is a dimension table of the star schema. As it shows above, the Product dimension needs two source files as the input. One is the 'ProductOrder.xlsx', and the other one is the 'Product.xlsx'.

In order to easily manage the margin, we need to calculate the cost first. Therefore, using Cost-Calculator step, we can access to the cost of each order directly. In this step, we multiply the Quantity with the PricePerItem to get the Cost. Then, using Product-Stream lookup step, we are able to retrieve the Group field from the Product file and merge it with the last step output.

Because one product is usually ordered many times, the cost of each order is always different. it is necessary to store the historic data preparing for the future analysis. Therefore, a slowly changing dimension is designed here. Using the 'Dimension lookup/update' step, we set up a Kimball Type II slowly changing, and linked it with a MySQL table called Cost.

Product_ID	version	date_from	date_to	id	Description	Cost	Date	Group
1	1	NULL	NULL	NULL	NULL	NULL	NULL	NULL
2	1	1900-01-01	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	15929.60...	2014-01-07	Lounge
3	2	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	10332.72	2014-01-26	Lounge
4	3	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	13489.93...	2014-05-27	Lounge
5	4	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	4487.400...	2014-07-08	Lounge
6	5	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	1085.840...	2015-01-17	Lounge
7	6	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	18518.23...	2015-08-01	Lounge
8	7	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	4629.559...	2015-08-05	Lounge
9	8	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	7822.36	2015-10-14	Lounge
10	9	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	9093.24	2016-01-22	Lounge
11	10	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	11131.38	2016-03-20	Lounge
12	11	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	9602.48	2016-08-17	Lounge
13	12	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	16059.32	2016-12-26	Lounge
14	13	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	2913.97	2017-01-16	Lounge
15	14	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	18683.69	2017-05-23	Lounge
16	15	2018-02-06	2018-02-06	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	19942.71...	2017-08-28	Lounge
17	1	1900-01-01	2018-02-06	100002	Ladies Essence Blouse Short Sleeve (LL-ES) (Black/White)	10045.69...	2014-02-03	Lounge
18	2	2018-02-06	2018-02-06	100002	Ladies Essence Blouse Short Sleeve (LL-ES) (Black/White)	1578.61	2014-06-12	Lounge
19	3	2018-02-06	2018-02-06	100002	Barron Mac Concealed (MCOJAC) (Black)	3014.88	2014-07-10	Lounge
20	4	2018-02-06	2018-02-06	100002	Barron Mac Concealed (MCOJAC) (Black)	17147.13	2014-08-23	Lounge
21	5	2018-02-06	2018-02-06	100002	Barron Mac Concealed (MCOJAC) (Black)	16509.68	2015-06-14	Lounge
22	6	2018-02-06	2018-02-06	100002	Barron Mac Concealed (MCOJAC) (Black)	21950.37...	2015-06-18	Lounge
23	7	2018-02-06	2018-02-06	100002	Barron Mac Concealed (MCOJAC) (Black)	1218.48	2015-07-06	Lounge
24	8	2018-02-06	2018-02-06	100002	Barron Mac Concealed (MCOJAC) (Black)	14621.76	2015-10-14	Lounge
25	9	2018-02-06	2018-02-06	100002	Barron Mac Concealed (MCOJAC) (Black)	13315.90...	2016-04-23	Lounge

Cost 1

Cost table

Lastly, we output the slowly changing dimension as a MySQL table named as Product table, which is shows below.

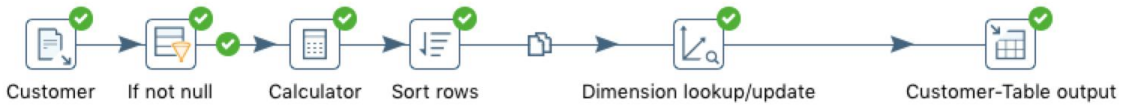
Prod_ID	Prod_Code	Description	Group	Cost	Date
2	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	15929.60…	2014-01-07
3	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	10332.72	2014-01-26
4	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	13489.93…	2014-05-27
5	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	4487.400…	2014-07-08
6	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	1085.840…	2015-01-17
7	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	18518.23…	2015-08-01
8	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	4629.559…	2015-08-05
9	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	7822.36	2015-10-14
10	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	9093.24	2016-01-22
11	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	11131.38	2016-03-20
12	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	9602.48	2016-08-17
13	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	16059.32	2016-12-26
14	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	2913.97	2017-01-16
15	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	18683.69	2017-05-23
16	100001	Ladies Essence Blouse Short Sleeve (LL-ES) (Grey/White)	Lounge	19942.71…	2017-08-28
17	100002	Ladies Essence Blouse Short Sleeve (LL-ES) (Black/White)	Lounge	10045.69…	2014-02-03
18	100002	Ladies Essence Blouse Short Sleeve (LL-ES) (Black/White)	Lounge	1578.61	2014-06-12
19	100002	Barron Mac Concealed (MCOJAC) (Black)	Lounge	3014.88	2014-07-10
20	100002	Barron Mac Concealed (MCOJAC) (Black)	Lounge	17147.13	2014-08-23
21	100002	Barron Mac Concealed (MCOJAC) (Black)	Lounge	16509.68	2015-06-14
22	100002	Barron Mac Concealed (MCOJAC) (Black)	Lounge	21950.37…	2015-06-18
23	100002	Barron Mac Concealed (MCOJAC) (Black)	Lounge	1218.48	2015-07-06
24	100002	Barron Mac Concealed (MCOJAC) (Black)	Lounge	14621.76	2015-10-14
25	100002	Barron Mac Concealed (MCOJAC) (Black)	Lounge	13315.90…	2016-04-23
26	100002	Barron Mac Concealed (MCOJAC) (Black)	Lounge	8194.400…	2016-04-23

Product 1

Product table

As we can see where the Prod_ID is 17 and Prod_Code is 100002, the Description field of this piece of data is same with the ones where Prod_Code is 100001. That is to say, the Prod_Code field of No.17 and 18 data is uncorrect. Here is a data quality issue that needs to be mended in the future.

2.3 Customer Transformation



Customer table is another dimension table of the data warehouse. In order to transform the source data into the target dimension table, the above six steps are used as it shows.

First, we extract the source data from Customer.xlsx file of the sale system. To reduce the space and avoid marking an ID to a null row, a 'If not null' step is used to delete the null rows. Next, as the stakeholder would like to analyse the business with the age group of the customers, the Calculator is the step to calculate each customer's age. And then, sort the rows with ascending CustCode along with the DateOfBirth.

Since the customers may change their postcode frequently, it is necessary to set up a 'Dimension lookup/update' step for slowly changing. We use CustCode as key fields to look up rows in dimension, and through auto increment field to create technical key. As a result, a new primary key is created called CustomerID, and a new table is created as Customer-slowlyChanging, which is shown as following figures.

CustomerID	version	date_from	date_to	CustCode	Name	Postcode	Date of Birth	Validuntil
1	1	NULL	NULL	NULL	NULL	NULL	NULL	NULL
2	1	1900-01-01	2200-01-01	C1002Carl	Dewi Chai Kagioglou	3005	1943-01-21	NULL
3	1	1900-01-01	2200-01-01	C1003Carl	Ludmilla Irawan Kaimak...	3173	1938-08-26	NULL
4	1	1900-01-01	2200-01-01	C1005Will	Eddie Lim Kan	3005	2000-08-25	NULL
5	1	1900-01-01	2200-01-01	C1006Carl	Melissa John Kang	3520	1979-02-14	NULL
6	1	1900-01-01	2200-01-01	C1007Will	Mei Tsun-To Kang	3072	1958-08-06	NULL
7	1	1900-01-01	2200-01-01	C1008Carl	Yeong Bernadette Kang	3053	1969-12-20	NULL
8	1	1900-01-01	2200-01-01	C1009Carl	Pauline Fei Kang	3068	1938-02-24	NULL
9	1	1900-01-01	2200-01-01	C100Carl	Xavier Heath Baohm	3041	1960-02-06	NULL
10	1	1900-01-01	2018-02-06	C1010Carl	Mohit Kumala Jia	3040	1955-12-11	NULL
11	2	2018-02-06	2018-02-06	C1010Carl	Mohit Kumala Jia	3056	1955-12-11	2012-10-17
12	1	1900-01-01	2200-01-01	C1012Will	Hsiao Leon Karen	3053	1984-04-07	NULL
13	1	1900-01-01	2200-01-01	C1013Carl	Bibin Yuen Kang	3051	1968-01-26	NULL
14	1	1900-01-01	2200-01-01	C1014Carl	Stewart Chak Karen	3173	1959-05-19	NULL
15	1	1900-01-01	2200-01-01	C1015Carl	Arun Christian Kang	3053	1956-02-22	NULL
16	1	1900-01-01	2200-01-01	C1016Carl	Kwan Kong Karen	3068	1949-10-20	NULL
17	1	1900-01-01	2200-01-01	C1017Carl	Greg J Karen	3053	1951-09-23	NULL
18	1	1900-01-01	2200-01-01	C1019Carl	Mingyu Koh Karen	3068	1942-11-03	NULL
19	1	1900-01-01	2200-01-01	C101Carl	Hanson Sze Baohm	3065	1960-11-30	NULL
20	1	1900-01-01	2200-01-01	C1021Will	Atlarelang Alberto Kassi...	3070	1992-10-24	NULL
21	1	1900-01-01	2200-01-01	C1024SthMelb	Nouras Arjun Kearney	3052	1962-09-15	NULL
22	1	1900-01-01	2200-01-01	C1025SthMelb	Liang Stuart Kearney	3040	1993-03-11	NULL
23	1	1900-01-01	2200-01-01	C1026Will	David Ann-Lin Kearney	3053	1957-06-28	NULL
24	1	1900-01-01	2200-01-01	C1027Carl	Hong Hoa Kearney	3053	1974-11-11	NULL
25	1	1900-01-01	2200-01-01	C1028Will	Hans Tong Kearney	3052	1938-05-27	NULL

Customer-slolyChanging 1

Customer-slowlyChanging table

CustomerID	CustCode	Name	Date of Birth	Age	Postcode	Validuntil
2	C1002Carl	Dewi Chai Kagioglou	1943-01-21	75	3005	NULL
3	C1003Carl	Ludmilla Irawan Kaimakamis	1938-08-26	80	3173	NULL
4	C1005Will	Eddie Lim Kan	2000-08-25	18	3005	NULL
5	C1006Carl	Melissa John Kang	1979-02-14	39	3520	NULL
6	C1007Will	Mei Tsun-To Kang	1958-08-06	60	3072	NULL
7	C1008Carl	Yeong Bernadette Kang	1969-12-20	49	3053	NULL
8	C1009Carl	Pauline Fei Kang	1938-02-24	80	3068	NULL
9	C100Carl	Xavier Heath Baohm	1960-02-06	58	3041	NULL
1314	C1010Carl	Mohit Kumala Jia	1955-12-11	63	3040	NULL
1315	C1010Carl	Mohit Kumala Jia	1955-12-11	63	3056	2012-10-17
12	C1012Will	Hsiao Leon Karen	1984-04-07	34	3053	NULL
13	C1013Carl	Bibin Yuen Kang	1968-01-26	50	3051	NULL
14	C1014Carl	Stewart Chak Karen	1959-05-19	59	3173	NULL
15	C1015Carl	Arun Christian Kang	1956-02-22	62	3053	NULL
16	C1016Carl	Kwan Kong Karen	1949-10-20	69	3068	NULL
17	C1017Carl	Greg J Karen	1951-09-23	67	3053	NULL
18	C1019Carl	Mingyu Koh Karen	1942-11-03	76	3068	NULL
19	C101Carl	Hanson Sze Baohm	1960-11-30	58	3065	NULL
20	C1021Will	Atlarelang Alberto Kassianos	1992-10-24	26	3070	NULL
21	C1024SthMelb	Nouras Arjun Kearney	1962-09-15	56	3052	NULL
22	C1025SthMelb	Liang Stuart Kearney	1993-03-11	25	3040	NULL
23	C1026Will	David Ann-Lin Kearney	1957-06-28	61	3053	NULL
24	C1027Carl	Hong Hoa Kearney	1974-11-11	44	3053	NULL
25	C1028Will	Hans Tong Kearney	1938-05-27	80	3052	NULL
26	C1029SthMelb	Nishant Sigit Kee	1994-04-23	24	3053	NULL

Customer 1

Customer table

2.4 Store Transformation

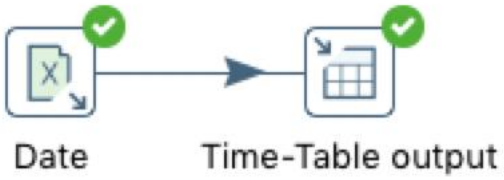
Regarding Store Table, there is only two steps involved in the transformation process. The first step is to input the table of store from the source data-Store.xlsx. Then the second step is to output the data of Store Table.

Store...	Description	Postc...
1	Carlton	3053
2	South Melbo...	3205
3	Williamstown	3016

Store 1

Store table

2.5 Date Transformation



The transformation for Date Table includes two simple steps: one is to input the related data from source file- Date.xlsx, another one is to output the data of Time Table.

TimeID	Date	DayOfWeek	Month	Quarter	Year	Day	Season
1	2014-01-01	Wednesday	January	1	2014	1	Summer
2	2014-01-02	Thursday	January	1	2014	2	Summer
3	2014-01-03	Friday	January	1	2014	3	Summer
4	2014-01-04	Saturday	January	1	2014	4	Summer
5	2014-01-05	Sunday	January	1	2014	5	Summer
6	2014-01-06	Monday	January	1	2014	6	Summer
7	2014-01-07	Tuesday	January	1	2014	7	Summer
8	2014-01-08	Wednesday	January	1	2014	8	Summer
9	2014-01-09	Thursday	January	1	2014	9	Summer
10	2014-01-10	Friday	January	1	2014	10	Summer
11	2014-01-11	Saturday	January	1	2014	11	Summer
12	2014-01-12	Sunday	January	1	2014	12	Summer
13	2014-01-13	Monday	January	1	2014	13	Summer
14	2014-01-14	Tuesday	January	1	2014	14	Summer
15	2014-01-15	Wednesday	January	1	2014	15	Summer
16	2014-01-16	Thursday	January	1	2014	16	Summer
17	2014-01-17	Friday	January	1	2014	17	Summer
18	2014-01-18	Saturday	January	1	2014	18	Summer
19	2014-01-19	Sunday	January	1	2014	19	Summer
20	2014-01-20	Monday	January	1	2014	20	Summer
21	2014-01-21	Tuesday	January	1	2014	21	Summer
22	2014-01-22	Wednesday	January	1	2014	22	Summer
23	2014-01-23	Thursday	January	1	2014	23	Summer

Time 1

Date table

3. Design of Data Warehouse

Based on the ETL process discussed above, we make redesign on our data warehouse and justify reasons for changes. This report uses the following figures to depict the original data warehouse design and the redesigned version respectively. Then this report demonstrates change issues regarding attributes from the perspective of the fact table and dimensional tables.

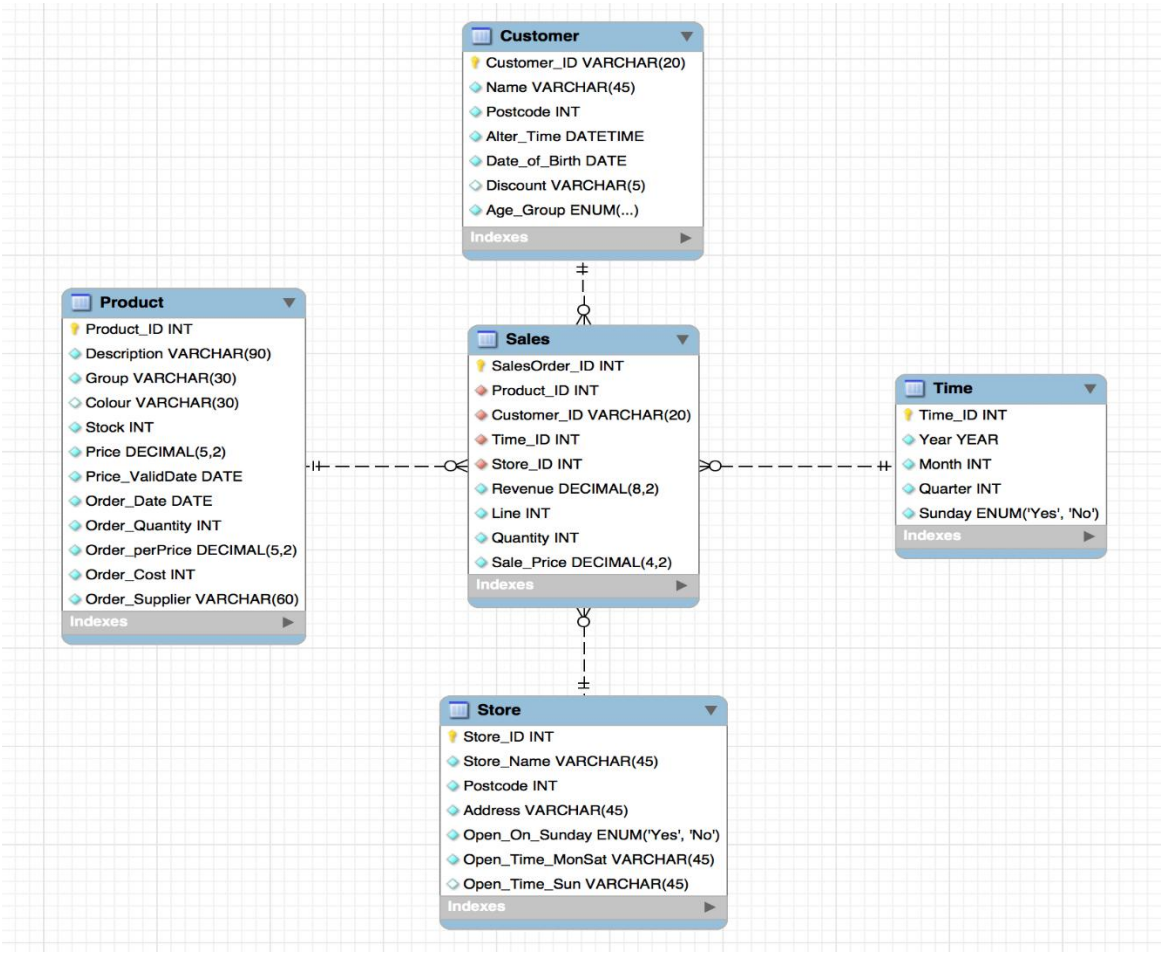


Figure 3.1 design of data warehouse (original version)

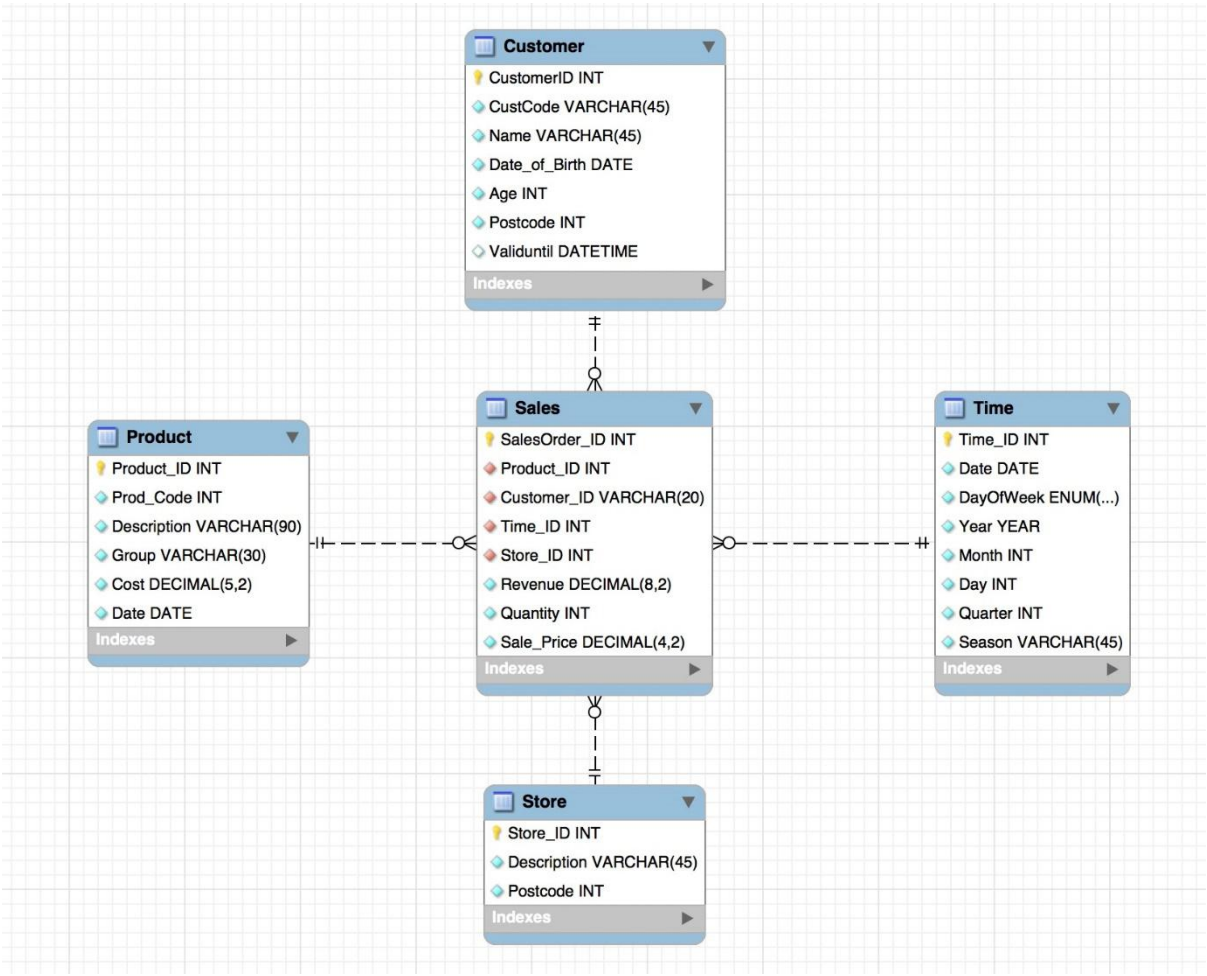


Figure 3.2 design of data warehouse (redesigned version)

3.1 Sales Table

In this table, the redesigned data warehouse removes line, which is used to identify unique products happened in same transaction by same customer. The reason is this attribute makes less contribution to both inventory system and sales operational system.

3.2 Product Table

In this table, the stock of each product for each inventory is removed, as it is more relevant to inventory system rather than sales system, which is the focus of this star schema. In addition, the aggregated quantity of each product, the supplier for each product, and the total cost of each product are also removed. The reasons for these three attributes' change is the existence of them makes product dimension table an operational system, which is an unclear and confusing design in sales system.

3.3 Time Table

Instead of using Sunday to indicate whether each sale happened in Sunday or not, the new model uses DayOfWeek to identify Sundays. The reason for this change is to avoid too much extraction and transformation jobs to do with this attribute. In detail, this field needs to be processed and analysed before employing it in the original design, while it could be used directly in the new model.

3.4 Customer Table

The only change in this table is to replace the group of age with the detailed age of each customers. To achieve more accurate divisions about age in the subsequent ETL process, specifying each customer's age could be a good choice.

3.5 Store Table

In this table, attributes about the detailed information of each store are removed, such as its name, address, opening hours in both weekdays and weekends. As the related information is not provided by source data, the existence of these attributes would cause data redundancy.

4. Data Dictionary

Sales Fact Table

Step	Description	Source
Sale	Extract source data	From the Sale.xlsx file
raw-Sale Table output	A pre-process transforming Excel file into MySQL table	From the Sale.xlsx file
raw-Sale Table input	Input the source data	From the raw-Sale table
Customer Table input		
CustomerID Stream lookup	Retrieve customerID field and add it to the Sale table	From Customer table
TimeID Stream lookup	Retrieve TimeID field and add it to the Sale table	From Time table
Calculator	Calculate revenue using UnitPrice multiplied by Unit	From the SaleItem.xlsx
ProductID Stream lookup	Get ProductID and add it to the Sale table	Using SaleID and ProductID from SaleItem.xlsx

Product Dimension

Step	Description	Source
ProductOrder	Extract source data as the input	ProductOrder.xlsx from the inventory system
Cost-Calculator	Calculate the cost of each	Calculated with CostPerItem

	order	multiplied by Quantity
Select values	Select ID, Description, Cost, and Date	From the last step
Product-Stream lookup	Retrieve Group field and add it to the ProductOder	From Produc.xlsx and the last step
Sort rows	Sort ProductCode in ascending order	From the last step
Dimension lookup/update	Create a new primary key-productID, to handle slowly changing cost for each product	From the last step
Table output	Output the new Product Dimension	From the last step

Customer Dimension

Step	**Description**	**Source**
Customer	Extract source data as the input	Customer.xlsx from the Order system
If-not-null	Filter data with non-null value	From the last step
Calculator	Calculate each customer's age	From the last step
Sort rows	Sort CustCode in ascending order	From the last step
Dimension lookup/update	Create a new primary key-CustomerID, to handle slowly changing postcode	From the last step
Customer-Table output	Output the Customer Table	From the last step

Store Dimension

Step	**Description**	**Source**
Store	Extract source data as the input	Store.xlsx from the Order system
Store-Table output	Output the Store Table	From the last step

Date Dimension

Step	**Description**	**Source**
Date	Extract source data as the input	Date.xlsx from the OLTP file
Time-Table output	Output the Time Table	From the last step

5.Appendix 1- Work Breakdown

Name	**Student ID**	**Contribution**

Ailin Zhang	874810	2.Design of the ETL Process 4.Data Dictionary
Nan Wang	853883	1.Executive Summary 3.Design of Data Warehouse 5.Appendix 1- WorkBreakdown 6. Appendix 2- Reference

Both members discussed the design of ETL process and redesign of data warehouse.

In detail, Ailin (874810) implements the ETL process for fact table and dimensional tables through Pentaho, and she completes most component of part 2 and part 4, which are design of the ETL Process and Data Dictionary.

While Nan (853883) accomplished part 1,3, 5 and 6, which are: executive summary, design of the Data Warehouse, work breakdown and the reference respectively. In addition, Nan completes partial components of part 4 and form editing for this report.

6.Appendix 2- Reference

Katariya, S. (2018). *Data warehouse Architecture and Process Flow. - Testingpool*. [online] Testingpool. Available at: http://testingpool.com/data-warehouse-architecture-and-etl-mechanism/ [Accessed 1 Feb. 2018].

Kimball Group. (2018). *Slowly Changing Dimensions, Part 2 - Kimball Group*. [online] Available at: https://www.kimballgroup.com/2008/09/slowly-changing-dimensions-part-2/ [Accessed 2 Feb. 2018].