ISYS90086 ASSIGNMENT 2 REPORT

2020/02/07

Student Name	Student Number
Hangyu Pan	1050937

Content

1. Executive Summary

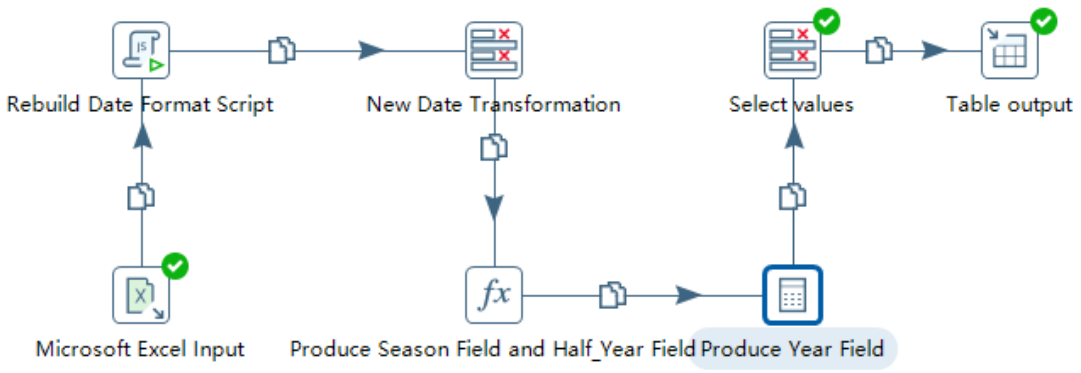
2. Design of the ETL Process

2.1 Extraction-Target Data and data sources

To address 5 areas for the business, it is necessary to identify the most profitable customer as the key customer, the product in maximum profit as the most profitable product, the market with maximum profit as the most profitable market as well as the sales agent with maximum sales as the key sales agent. As for time periods, it is depending on the period features for comparison such as season, month and year etc. Basically, the profitable time period does have maximum profit.

Based on the strategic information required, it can be assumed that the data extraction is unnecessary to be developed for real time, as addressing these 5 business concerns cannot be reflected in a short period. Therefore, it will implement deferred data extraction. Assuming that the reporting period for these 5 areas is monthly or weekly, comparing files with time stamped is significant for each extraction especially for the customer information and sales agent information. The capture can be addressed based on date and time stamp.

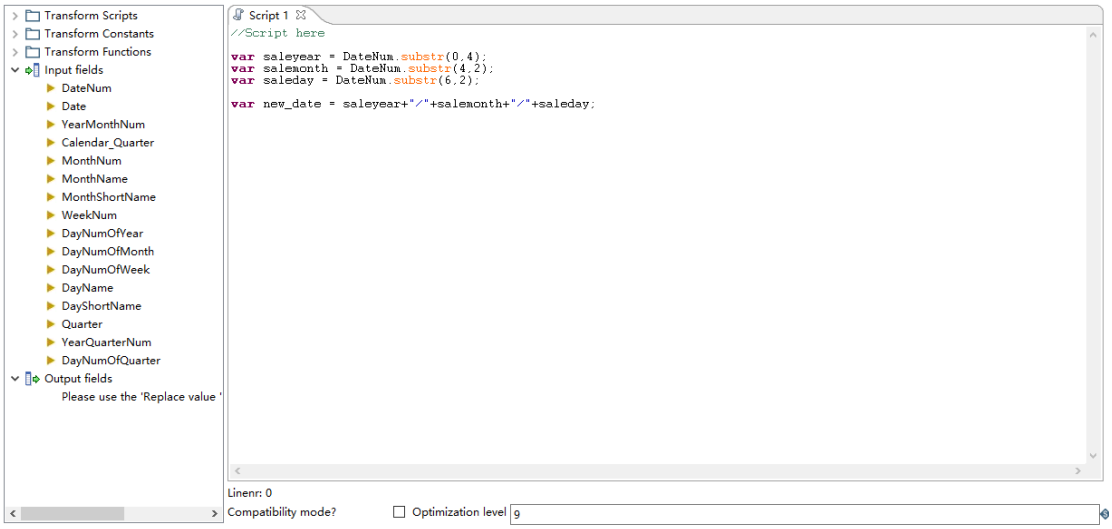
2.2.1 Date Transformation



In the date transformation, it contains 6 steps. The first step is to input Date.xlsx.

#	Name	Type	Length	Precision	Trim type	Repeat	Format	Currency	Decimal	Grouping
1	DateNum	String	-1	-1	none	N				
2	Date	Date	-1	-1	none	N	yyyy-MM-dd			
3	YearMonthNum	Integer	-1	-1	none	N				
4	Calendar_Quarter	String	-1	-1	none	N				
5	MonthNum	Integer	-1	-1	none	N				
6	MonthName	String	-1	-1	none	N				
7	MonthShortName	String	-1	-1	none	N				
8	WeekNum	Integer	-1	-1	none	N				
9	DayNumOfYear	Integer	-1	-1	none	N				
10	DayNumOfMonth	Integer	-1	-1	none	N				
11	DayNumOfWeek	Integer	-1	-1	none	N				
12	DayName	String	-1	-1	none	N				
13	DayShortName	String	-1	-1	none	N				
14	Quarter	Integer	-1	-1	none	N				
15	YearQuarterNum	Integer	-1	-1	none	N				
16	DayNumOfQuarter	Integer	-1	-1	none	N				

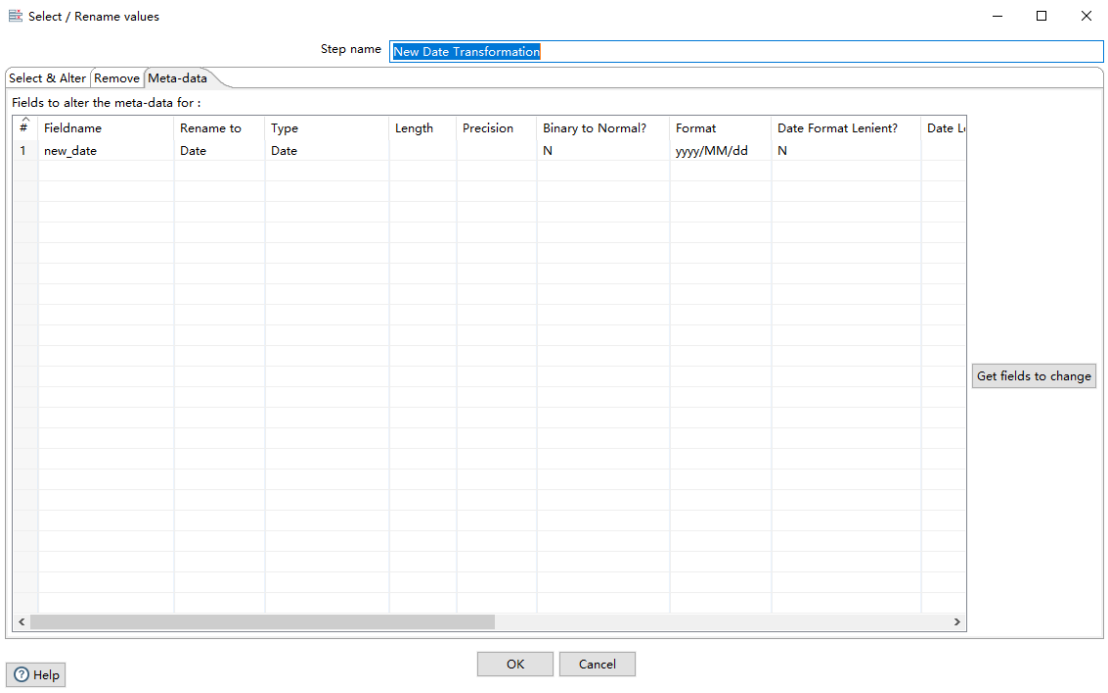
After date excel input, it is going to reformat date. It contains 2 steps. Through modified Java Script Value, it is to rebuild a new date filed by DateNum.



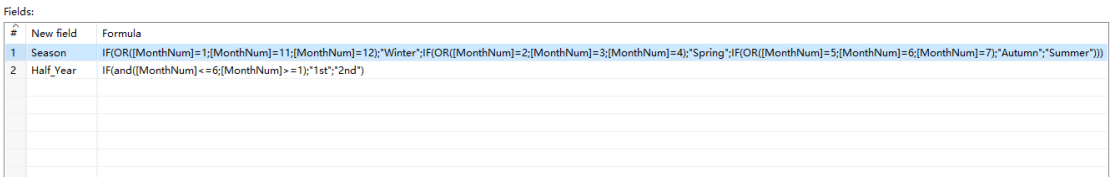
Fields

#	Fieldname	Rename to	Type	Length	Precision	Replace value 'Fieldname' or 'Rename to'
1	new_date		String			N

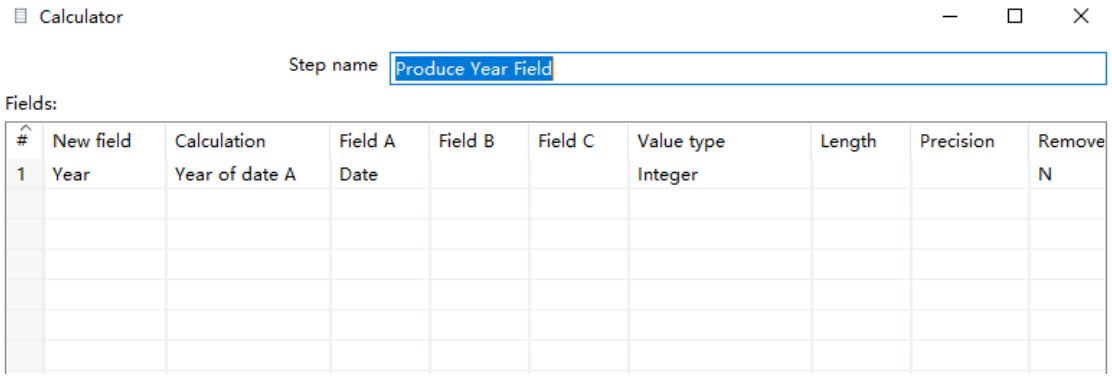
Then, the step of new date transformation is going to use new date to replace Date fields. The date format is to be yyyy/MM/dd instead of yyyy-MM-dd.



For next part, it is going to extract some of features from the date. Through the step of formula, it is to generate two derived values from date which are season and half_year identification.



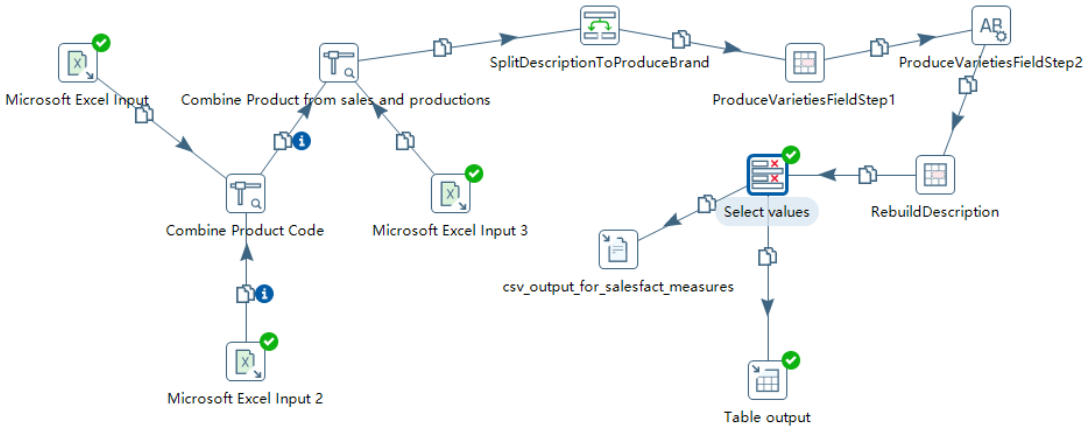
Next, the step of calculator is to produce year field.



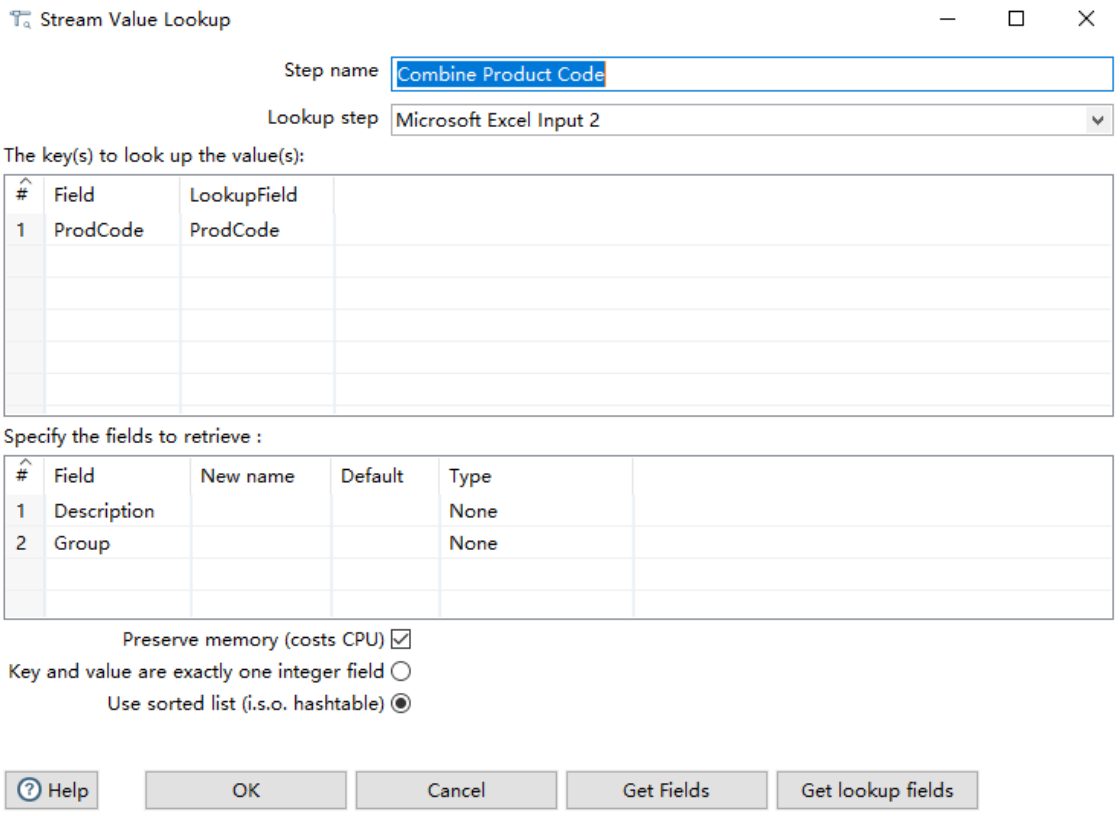
Finally, it is to select required value for the dimension and get table output. For the most of time, date dimension is no need to be updated. It can insert date dimension one off within the range up to the project. Here is the part of output after connecting to data warehouse:

Date	Day_Of_Week	MonthName	Week_Number	Half_Year	Year	Season_Name	Quarter	DateNumber
1991-01-01	Tuesday	Jan	1	1st	1991	Winter	1	19910101
1991-01-02	Wednesday	Jan	1	1st	1991	Winter	1	19910102
1991-01-03	Thursday	Jan	1	1st	1991	Winter	1	19910103
1991-01-04	Friday	Jan	1	1st	1991	Winter	1	19910104
1991-01-05	Saturday	Jan	1	1st	1991	Winter	1	19910105
1991-01-06	Sunday	Jan	1	1st	1991	Winter	1	19910106
1991-01-07	Monday	Jan	2	1st	1991	Winter	1	19910107
1991-01-08	Tuesday	Jan	2	1st	1991	Winter	1	19910108
1991-01-09	Wednesday	Jan	2	1st	1991	Winter	1	19910109
1991-01-10	Thursday	Jan	2	1st	1991	Winter	1	19910110
1991-01-11	Friday	Jan	2	1st	1991	Winter	1	19910111
1991-01-12	Saturday	Jan	2	1st	1991	Winter	1	19910112
1991-01-13	Sunday	Jan	2	1st	1991	Winter	1	19910113
1991-01-14	Monday	Jan	3	1st	1991	Winter	1	19910114
1991-01-15	Tuesday	Jan	3	1st	1991	Winter	1	19910115
1991-01-16	Wednesday	Jan	3	1st	1991	Winter	1	19910116
1991-01-17	Thursday	Jan	3	1st	1991	Winter	1	19910117
1991-01-18	Friday	Jan	3	1st	1991	Winter	1	19910118
1991-01-19	Saturday	Jan	3	1st	1991	Winter	1	19910119

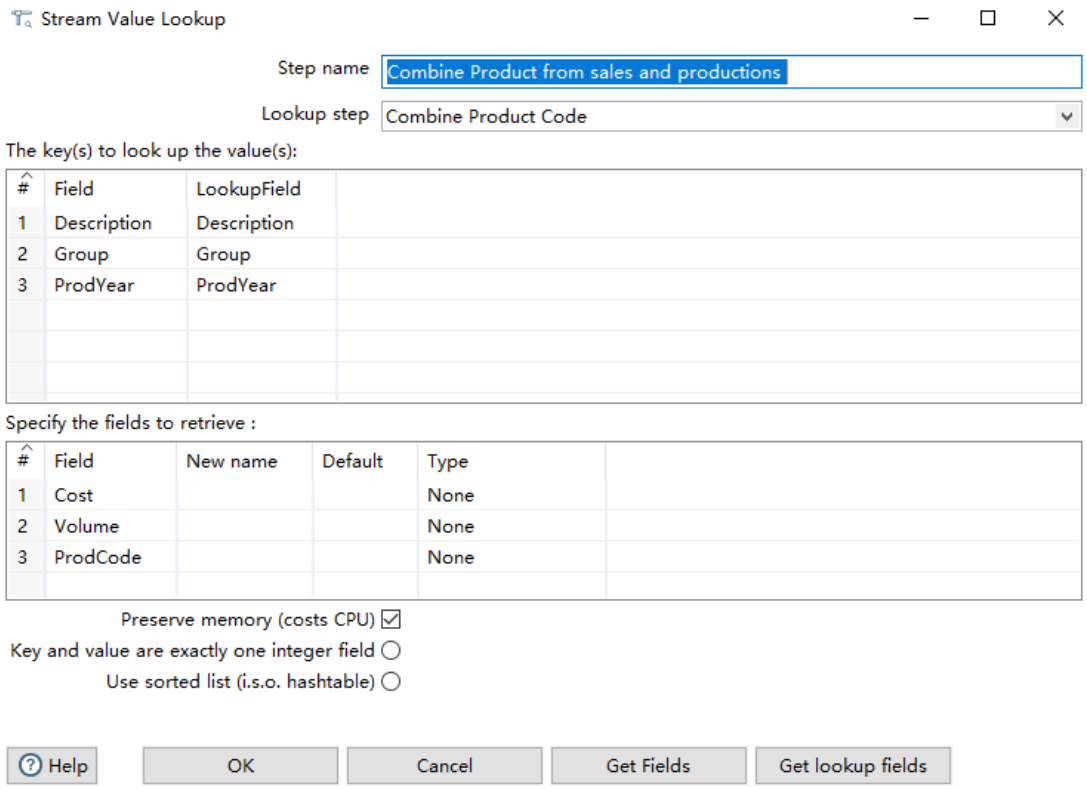
2.2.2 Product Transformation



For product transformation, it does contain 12 steps. Firstly, the data of product from production system has two types of information. The step of combine product code is to map each ProductionID with its own Prod Code to get the production detail. The first input is for the production history and second input is for the product detail in the production system.



Then, it is required to combine product detail to grab the price field. Through the step of combine product from sales and productions, it can grab the basic price for the product. The third input file is product detail according to the sales system. To map the field, it puts description, groups and prodyear as lookup field and get output for product table with new fields of cost, volume as stock on hand and Prodcode.



To grab the brand and varieties detail from the description, it contains several steps which are SplitDescription and ConcatString to grab Varieties as well as rebuild description. Finally, it can grab temporary output table.

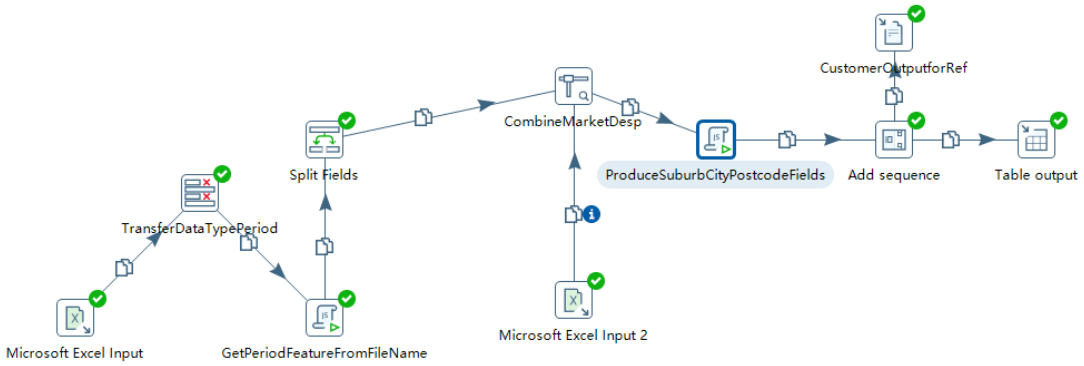
ProductKey	ProdCode	Brand_Name	Varieties	Description	ProdYear	Group	Unit_Price	Unit_Cost	Stock_On_Hand
1	1	Bellarine	Pinot Grigio	Bellarine Pinot Grigio	2014	White	163	111	11587
2	2	Bellarine	Pinot Noir	Bellarine Pinot Noir	2014	Red	113	67	11250
3	3	Downunder	Merlot	Downunder Merlot	2014	Red	127	79	10959
4	4	Downunder	Pinot Grigio	Downunder Pinot Grigio	2014	White	79	54	9896
5	5	Downunder	Pinot Noir	Downunder Pinot Noir	2014	Red	100	83	13850
6	6	Overhill	Merlot	Overhill Merlot	2014	Red	135	82	9565
7	7	Overhill	Pinot Noir	Overhill Pinot Noir	2014	Red	98	73	12594
8	1	Bellarine	Pinot Grigio	Bellarine Pinot Grigio	2014	White	139	111	11587
9	2	Bellarine	Pinot Noir	Bellarine Pinot Noir	2014	Red	106	67	11250
10	3	Downunder	Merlot	Downunder Merlot	2014	Red	111	79	10959
11	4	Downunder	Pinot Grigio	Downunder Pinot Grigio	2014	White	85	54	9896
12	5	Downunder	Pinot Noir	Downunder Pinot Noir	2014	Red	116	83	13850
13	6	Overhill	Merlot	Overhill Merlot	2014	Red	125	82	9565
14	7	Overhill	Pinot Noir	Overhill Pinot Noir	2014	Red	97	73	12594
15	1	Bellarine	Pinot Grigio	Bellarine Pinot Grigio	2015	White	177	112	10176

For the output, it generates two files. One csv_output is for the reference for salesfact transformation to calculate measures. The other is to integrate data warehouse. The part of output is:

Product ID	Production_Code	Brand_Name	Varieties	Description	Group	Unit_Cost(Doz)	Unit_Price(Doz)	Stock_On_Hand	ProdYear
1	1	Bellarine	Pinot Grigio	Bellarine Pinot Grigio	White	111.00	163.00	11587	2014
2	2	Bellarine	Pinot Noir	Bellarine Pinot Noir	Red	67.00	113.00	11250	2014
3	3	Downunder	Merlot	Downunder Merlot	Red	79.00	127.00	10959	2014
4	4	Downunder	Pinot Grigio	Downunder Pinot Grigio	White	54.00	79.00	9896	2014
5	5	Downunder	Pinot Noir	Downunder Pinot Noir	Red	83.00	100.00	13850	2014
6	6	Overhill	Merlot	Overhill Merlot	Red	82.00	135.00	9565	2014
7	7	Overhill	Pinot Noir	Overhill Pinot Noir	Red	73.00	98.00	12594	2014
8	1	Bellarine	Pinot Grigio	Bellarine Pinot Grigio	White	111.00	139.00	11587	2014
9	2	Bellarine	Pinot Noir	Bellarine Pinot Noir	Red	67.00	106.00	11250	2014
10	3	Downunder	Merlot	Downunder Merlot	Red	79.00	111.00	10959	2014
11	4	Downunder	Pinot Grigio	Downunder Pinot Grigio	White	54.00	85.00	9896	2014
12	5	Downunder	Pinot Noir	Downunder Pinot Noir	Red	83.00	116.00	13850	2014
13	6	Overhill	Merlot	Overhill Merlot	Red	82.00	125.00	9565	2014
14	7	Overhill	Pinot Noir	Overhill Pinot Noir	Red	73.00	97.00	12594	2014
15	1	Bellarine	Pinot Grigio	Bellarine Pinot Grigio	White	112.00	177.00	10176	2015
16	2	Bellarine	Pinot Noir	Bellarine Pinot Noir	Red	74.00	117.00	11417	2015
17	3	Downunder	Merlot	Downunder Merlot	Red	86.00	127.00	11365	2015
18	4	Downunder	Pinot Grigio	Downunder Pinot Grigio	White	57.00	87.00	11638	2015
19	5	Downunder	Pinot Noir	Downunder Pinot Noir	Red	95.00	129.00	12628	2015
20	6	Overhill	Merlot	Overhill Merlot	Red	90.00	170.00	9841	2015
21	7	Overhill	Pinot Noir	Overhill Pinot Noir	Red	77.00	125.00	10469	2015
22	1	Bellarine	Pinot Grigio	Bellarine Pinot Grigio	White	121.00	151.00	6614	2016
23	2	Bellarine	Pinot Noir	Bellarine Pinot Noir	Red	81.00	136.00	7527	2016
24	3	Downunder	Merlot	Downunder Merlot	Red	90.00	139.00	6034	2016
25	4	Downunder	Pinot Grigio	Downunder Pinot Grigio	White	60.00	95.00	6051	2016
26	5	Downunder	Pinot Noir	Downunder Pinot Noir	Red	99.00	123.00	6540	2016
27	6	Overhill	Merlot	Overhill Merlot	Red	97.00	149.00	6933	2016
28	7	Overhill	Pinot Noir	Overhill Pinot Noir	Red	83.00	114.00	5294	2016
29	1	Bellarine	Pinot Grigio	Bellarine Pinot Grigio	White	80.00	167.00	1120	2017
30	2	Bellarine	Pinot Noir	Bellarine Pinot Noir	Red	45.00	140.00	1090	2017
31	3	Downunder	Merlot	Downunder Merlot	Red	65.00	152.00	1349	2017
32	4	Downunder	Pinot Grigio	Downunder Pinot Grigio	White	41.00	97.00	423	2017
33	5	Downunder	Pinot Noir	Downunder Pinot Noir	Red	60.00	152.00	1422	2017
34	6	Overhill	Merlot	Overhill Merlot	Red	58.00	158.00	1187	2017
35	7	Overhill	Pinot Noir	Overhill Pinot Noir	Red	50.00	114.00	700	2017
36	1	Bellarine	Pinot Grigio	Bellarine Pinot Grigio	White	84.00	164.00	3700	2018
37	2	Bellarine	Pinot Noir	Bellarine Pinot Noir	Red	51.00	158.00	3243	2018
38	3	Downunder	Merlot	Downunder Merlot	Red	63.00	158.00	4655	2018
39	4	Downunder	Pinot Grigio	Downunder Pinot Grigio	White	40.00	94.00	4207	2018
40	5	Downunder	Pinot Noir	Downunder Pinot Noir	Red	67.00	145.00	4737	2018
41	6	Overhill	Merlot	Overhill Merlot	Red	66.00	153.00	5313	2018
42	7	Overhill	Pinot Noir	Overhill Pinot Noir	Red	54.00	112.00	5298	2018

2.2.3 Customer Transformation

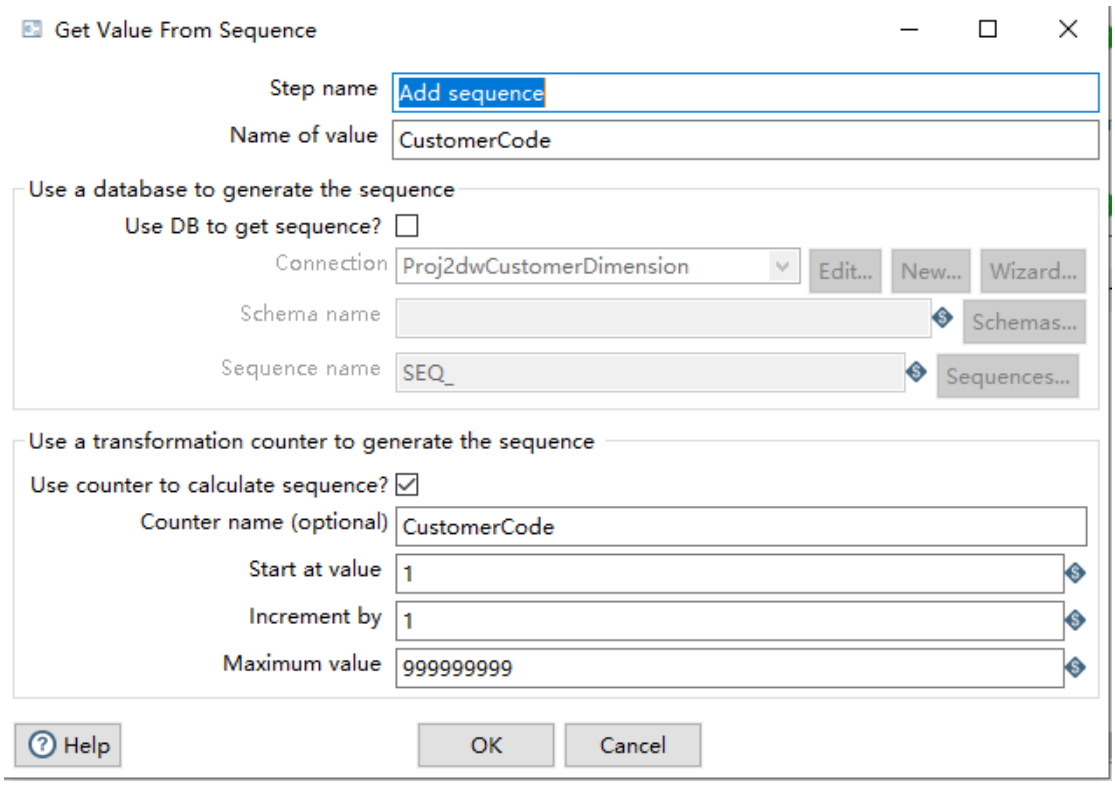
For customer dimension, it can be found that the customer has monthly data updates from the transaction system such as address and markets as well as its own derived values. Basically, the address and derived values like suburbs, city and postcode will be processed as type 2 for SCDs as same as the market value, as these kinds of data is changed for a period of time. Therefore, to use type 2 insertion does avoid the impact of new address changes on sales. For instance, if the sales were recorded with the customer from Victoria but he changes to Victoria after Feb 19, it is necessary to identify the sale before Mar 19 is in the Victoria market and sale after Feb 19 is for the international market. Type 1 update for SCDs is for the name of customer which is to correct information up to date.



In the transformation step, there are multiple data input from Jan 18, Feb 19 and December 19. The steps of transferdatatypeperiod and GetPeriodFeatureFromFileName is to extract file name time stamp for dimension update as well as produce new field of period to represent the data period.

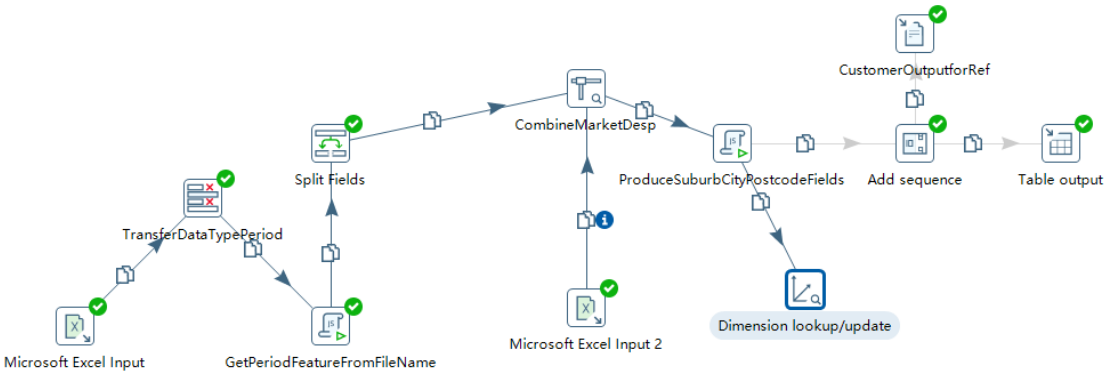
Then, to combine the market description, it does include steps of CombineMarektDesp, Input 2. New fields of suburb, city and postcode will be produced through the java script.

In addition, the new customer code will be used as surrogate key for the customer dimension through step of add sequence.



There will be two outputs. One is for the reference for salesfact input. The other is to integrate data to data warehouse.

For the further update, it is necessary to set up dimension lookup to identify the types and update action.



Dimension Lookup / Update

Step name: Dimension lookup/update

Update the dimension? ☑

Connection: Proj2dwCustomerDimension [Edit...] [New...] [Wizard...]

Target schema: [Browse...]

Target table: customer_dimension [Browse...]

Commit size: 100

Enable the cache? ☑

Pre-load the cache? ☐

Cache size in rows (0 = cache all): 5000

Keys | Fields

Key fields (to look up row in dimension):

#	Dimension field	Field in stream
1	Customer_ID	Cust ID

Technical key field: CustomerCode — New name:

Creation of technical key

- ○ Use table maximum + 1
- ○ Use sequence
- ◉ Use auto increment field

Version field: version

Stream Datefield:

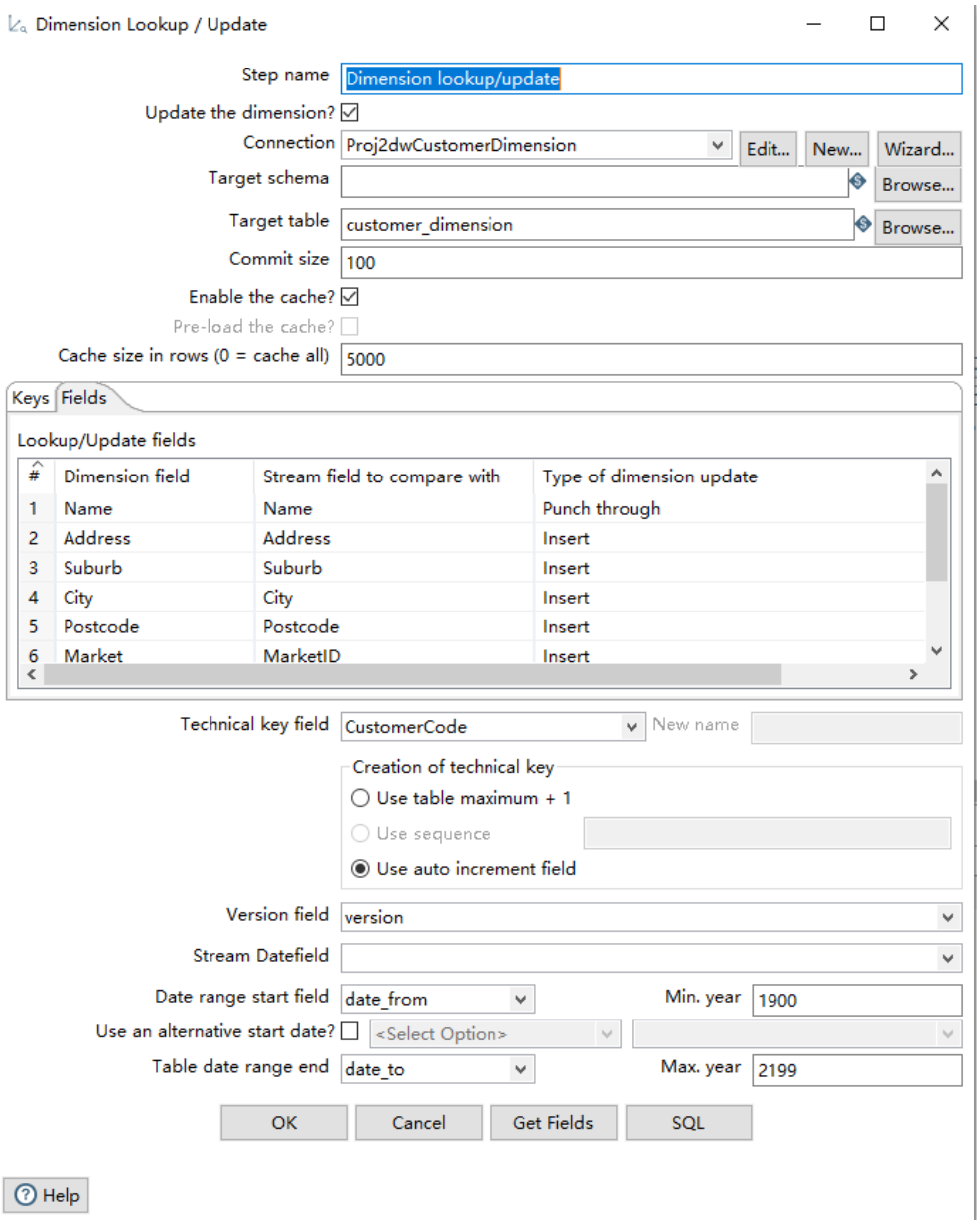
Date range start field: date_from — Min. year: 1900

Use an alternative start date? ☐ <Select Option>

Table date range end: date_to — Max. year: 2199

[OK] [Cancel] [Get Fields] [SQL]

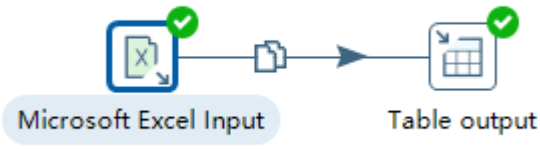
[Help]



Here is part of output for the first input in the data warehouse:

CustomerCode	Customer_ID	Period	Customer_Name	Address	Suburb	City	Postcode	Market	Market Description
36	13	Jan 2018	The Wine Rep	Smith St.,Collingwood,Melbourne 3066	Collingwood	Melbourne	3066	Vic	Victoria
37	14	Jan 2018	Gangemis Fine Wine and Foods	Hay Street,West Perth,Perth 6005	West Perth	Perth	6005	Aus	Rest of Australia
38	15	Jan 2018	Aussie Boutique Wines	Springvale Rd.,Springvale South,Melbourne 3172	Springvale South	Melbourne	3172	Vic	Victoria
39	16	Jan 2018	Armadale Cellars	High Street,Armadale,Melbourne 3143	Armadale	Melbourne	3143	Vic	Victoria
4	5	Feb 2019	Merchant's Lair	Nepean Highway,Mentone,Melbourne 3194	Mentone	Melbourne	3194	Vic	Victoria
40	17	Jan 2018	Dande Upmarket Wines	Pikkles St.,Dandenong,Melbourne 3175	Dandenong	Melbourne	3175	Vic	Victoria
41	18	Jan 2018	Family Wines Direct	Miamup Road,Cowaramup,Perth 6284	Cowaramup	Perth	6284	Aus	Rest of Australia
42	19	Jan 2018	Fine Wine Merchant	Mount Eliza Way,Mt Eliza,Melbourne 3930	Mt Eliza	Melbourne	3930	Vic	Victoria
43	1	Dec 2019	Zelas Wines	Archway Road,London ,London N6 5AX	London	London	N6 5AX	Int	International
44	2	Dec 2019	Oz Wines	Little St.,Richmond,Melbourne 3121	Richmond	Melbourne	3121	Vic	Victoria
45	4	Dec 2019	The Sussex Wine Company	Birdham Road,Chichester,West Sussex PO20 7DU	Chichester	West Sus...	PO20 7DU	Int	International
46	5	Dec 2019	Merchant's Lair	Nepean Highway,Mentone,Melbourne 3194	Mentone	Melbourne	3194	Vic	Victoria
47	6	Dec 2019	Australia Wines Direct	High St.,Stourbridge,West Midlands DY8 1TA	Stourbridge	West Midl...	DY8 1TA	Int	International
48	10	Dec 2019	La Cantina at Mercato	Lower North East Rd,Campbelltown,Adelaide 5074	Campbelltown	Adelaide	5074	Aus	Rest of Australia
49	11	Dec 2019	T & A Wines	Station Way,Brandon,Suffolk IP27 0BH	Brandon	Suffolk	IP27 0BH	Int	International
5	6	Feb 2019	Australia Wines Direct	High St.,Stourbridge,West Midlands DY8 1TA	Stourbridge	West Midl...	DY8 1TA	Int	International
50	12	Dec 2019	Acme Wine Imports	High St,Fullham,London SW1A 1LZ	Fullham	London	SW1A 1LZ	Int	International
51	13	Dec 2019	The Wine Rep	Smith St.,Collingwood,Melbourne 3066	Collingwood	Melbourne	3066	Vic	Victoria
52	14	Dec 2019	Gangemis Fine Wine and Foods	Hay Street,West Perth,Perth 6005	West Perth	Perth	6005	Aus	Rest of Australia
53	15	Dec 2019	Aussie Boutique Wines	Springvale Rd.,Springvale South,Melbourne 3172	Springvale South	Melbourne	3172	Vic	Victoria
54	18	Dec 2019	Family Wines Direct	Miamup Road,Cowaramup,Perth 6284	Cowaramup	Perth	6284	Aus	Rest of Australia
55	19	Dec 2019	Fine Wine Merchant	Mount Eliza Way,Mt Eliza,Melbourne 3930	Mt Eliza	Melbourne	3930	Vic	Victoria
56	3	Dec 2019	London Wines	King St.,London ,London SW1A 1LZ	London	London	SW1A 1LZ	Int	International
57	9	Dec 2019	Justerini & Brooks	The Strand,London ,London SW1A 1LZ	London	London	SW1A 1LZ	Int	International
58	8	Dec 2019	The Wine Club	James St,Nth. Melb,Melbourne 3051	Nth. Melb	Melbourne	3051	Vic	Victoria
59	20	Dec 2019	The Wine Room	Tankerton Road,Whitstable ,Whitstable CT5 2AJ	Whitstable	Whitstable	CT5 2AJ	Int	International
6	7	Feb 2019	Prestige Wines	Lygon St.,Carlton,Melbourne 3053	Carlton	Melbourne	3053	Vic	Victoria
60	21	Dec 2019	Galah Wine	Sturt Hwy,Ashton,Adelaide 5137	Ashton	Adelaide	5137	Aus	Rest of Australia
61	22	Dec 2019	Leon Stolarski Fine Wines	Nottingham Road,Hucknall,Nottingham NG15 7QE	Hucknall	Nottingham	NG15 7QE	Int	International
62	23	Dec 2019	Liquor Barons	Cambridge Street,Wembley,Perth 6014	Wembley	Perth	6014	Aus	Rest of Australia
63	24	Dec 2019	Merricks Wine Merchants	Frankston - Flinders Road,Merricks,Melbourne 3...	Merricks	Melbourne	3916	Vic	Victoria

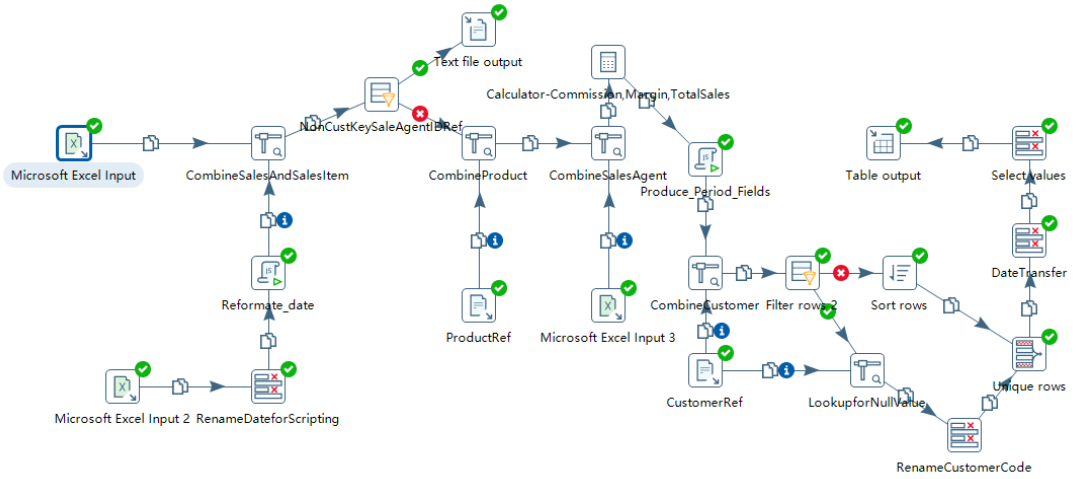
2.2.4 Sales Agent Transformation



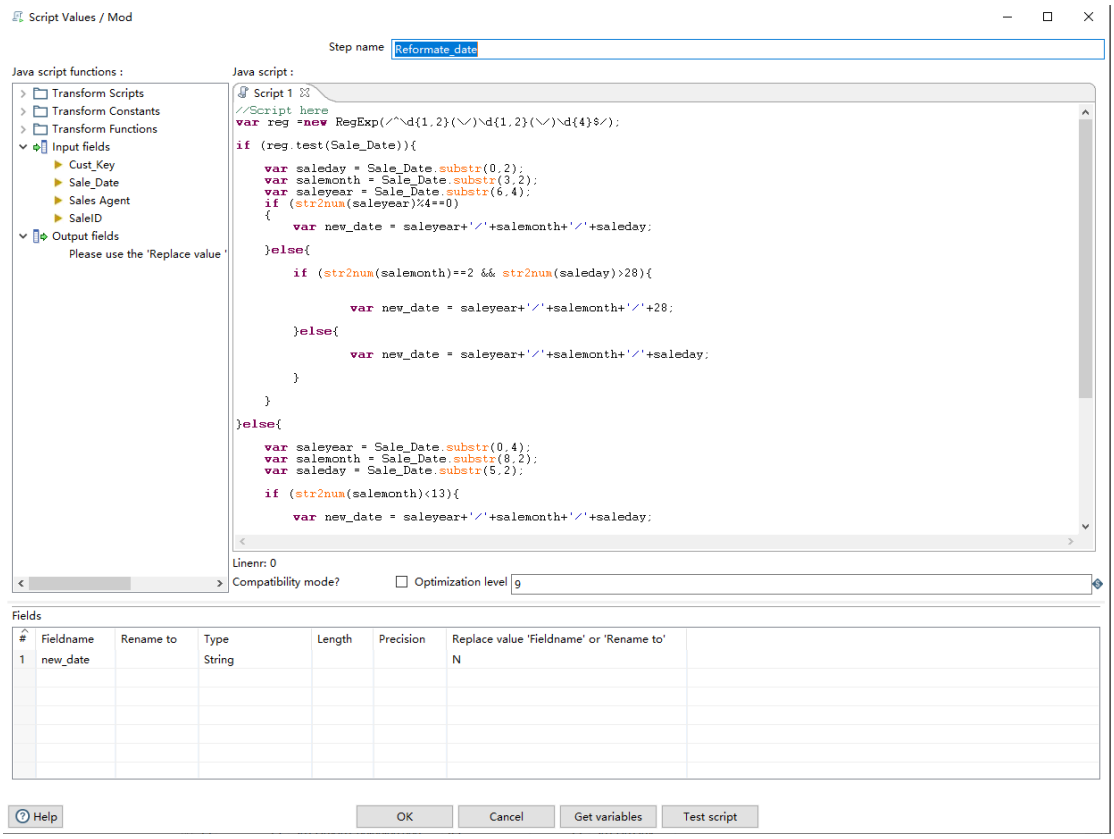
In the sales agent transformation, there are two steps. The input is Sales Agent.xlsx. Here is the part of output:

Employee_ID	Name	Commission_Rate
B1	Supradeek Densiman	0.2
B2	Arit Arubne	0.12
B3	Flame Blower	0.07
B4	Michelle Nguyen	0.07
D1	Hi Min Chow	0.19
D2	Peter Jones	0.08
D3	Aimee Concroan	0.07
D4	Jan Kennedy	0.04
M1	Alice McPherson	0.09
M2	Pjan Ling	0.03
S1	Willy Wonka	0.18
S2	Quin Tan	0.05
NULL	NULL	NULL

2.2.5 Sales Fact Transformation

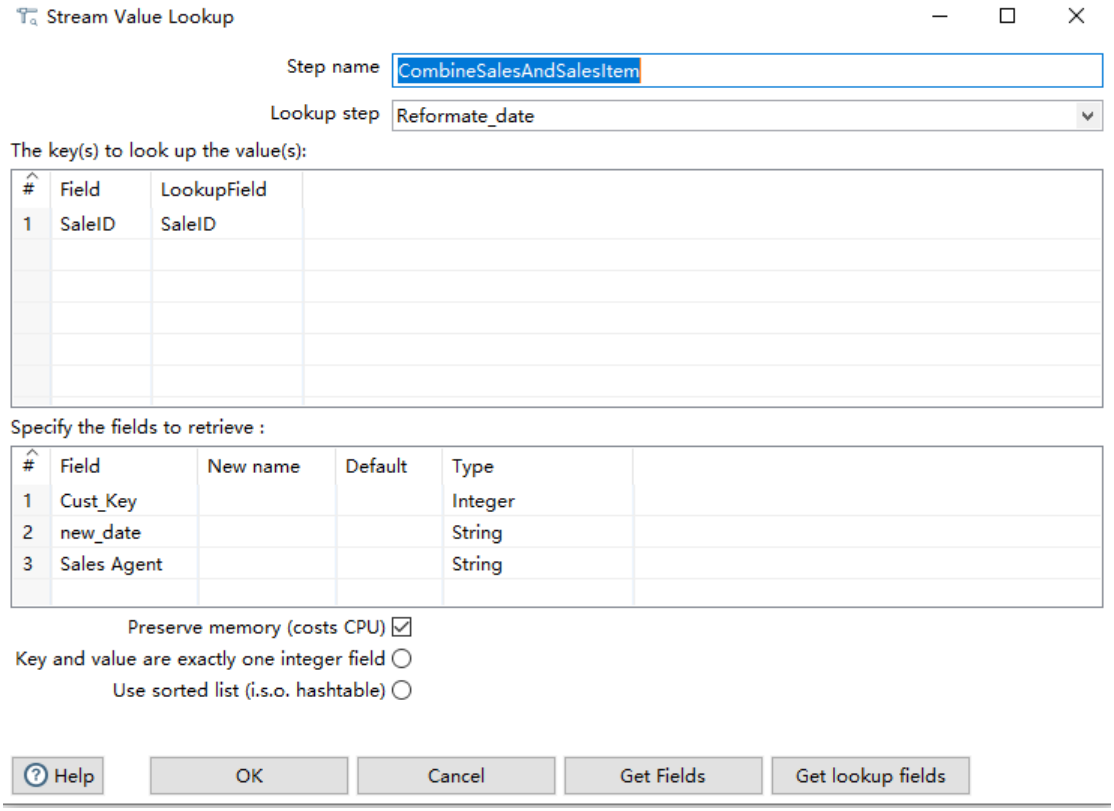


In the Sales Fact Transformation, it does contain 23 steps to deal with various situation. For inputs, the critical issue is sales date. Because of raw data from various system with its own time setting, it does lead the hardship to recognise the correct date. Therefore, the date field has to be pre-processed through script after getting input in string type.



There are two major issues in date. The first one is that there are sale dates wrongly recorded in 29/2/2017 which does not exist. It is corrected to be 28/2/2017. Secondly, some sale date is wrongly recognised by excel. For instance, the date of Feb 1st 2017, it is wrongly recognized as Jan 2nd 2017.

Firstly, it is to combine sales with sales item. In this procedure, it can find each dimension reference for sales based on the sale_id.

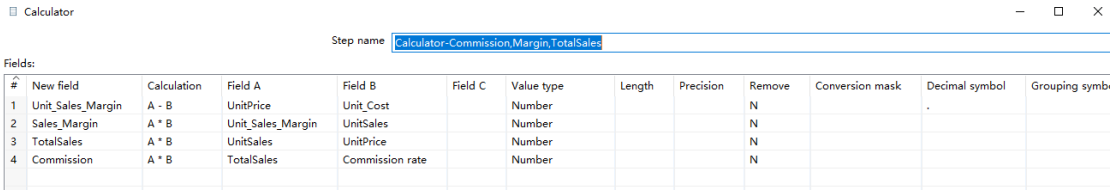


Additionally, there are few sales without the dimension references. It is going to filter it out and make further analysis.

#	SaleID	LineID	Prod_Key	UnitSales	UnitPrice	Cust_Key	new_date	Sales Agent
1	2012	1	27	101	136.00	<null>	<null>	<null>
2	2013	1	25	53	83.00	<null>	<null>	<null>
3	2014	1	16	88	118.00	<null>	<null>	<null>

Up to null value output, it does not have any ways to help them identify the dimension reference. Therefore, they will be ignored.

In the next procedures, it is to combine the product table to get reference product key for the product and to combine sales agent table to get reference employee id for the commission rate. After that, it is going to going to calculate sales margin and commission for line of sales.



Then, for next procedures to grab the customer dimension reference, it is necessary to allocate sales date and customer key to the customer reference table to grab the new

CustomerCode. Basically, it does link with Period and customer ID. Therefore, it is necessary to grab the period fields through scripts.

```
//Script here
var temp=replace(new_date,"/","");
var date_compare=parseInt(temp);
if (date_compare<20190201){

    var Period = "Jan 2018";

}else if(date_compare<20191201){

    var Period = "Feb 2019";

}else{

    var Period = "Dec 2019"

}
```

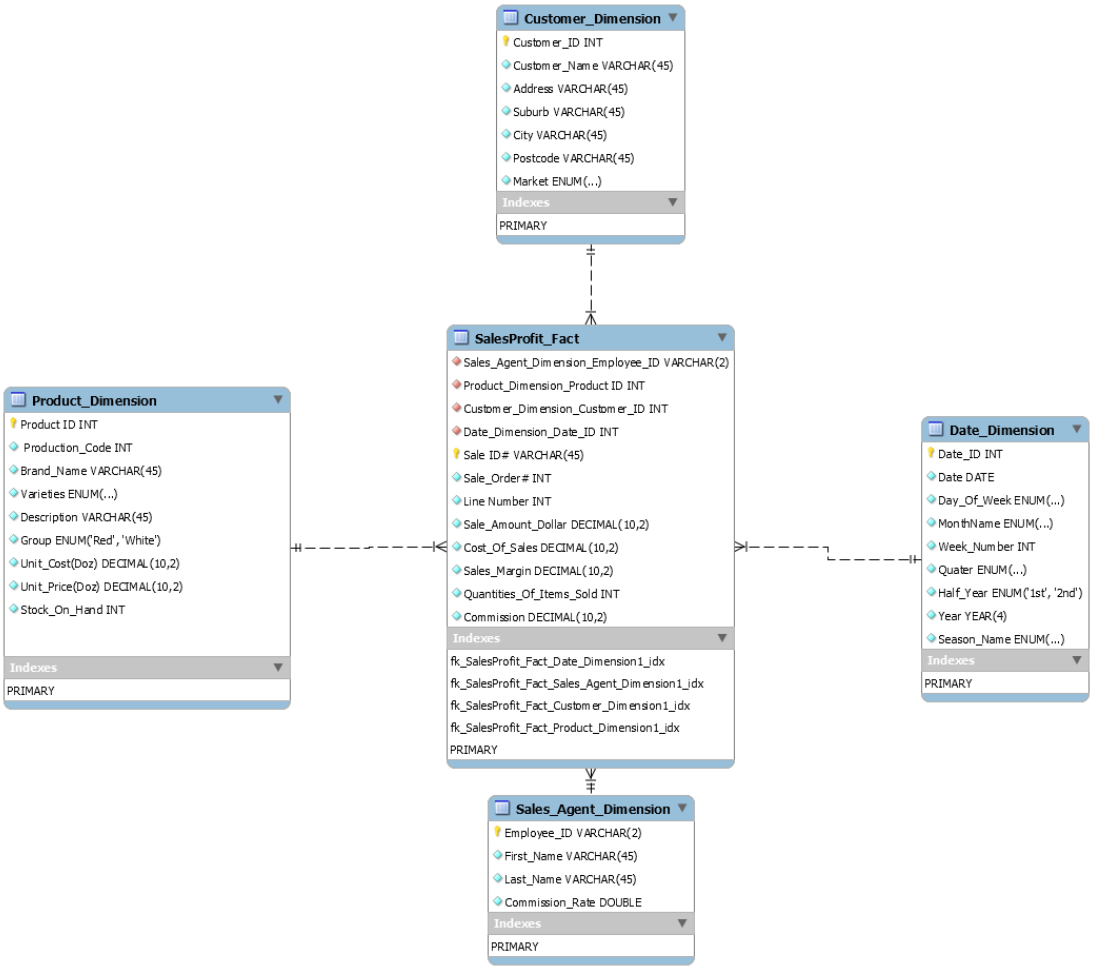
Up to the null value without correct reference period and customer key, it can use the customer key to map the customer ID to find the new customer code. All of sales can be allocated with its own CustomerCode.

Finally, it is to integrate the salesfact to data warehouse. Here is the part of output of salesprofit_fact:

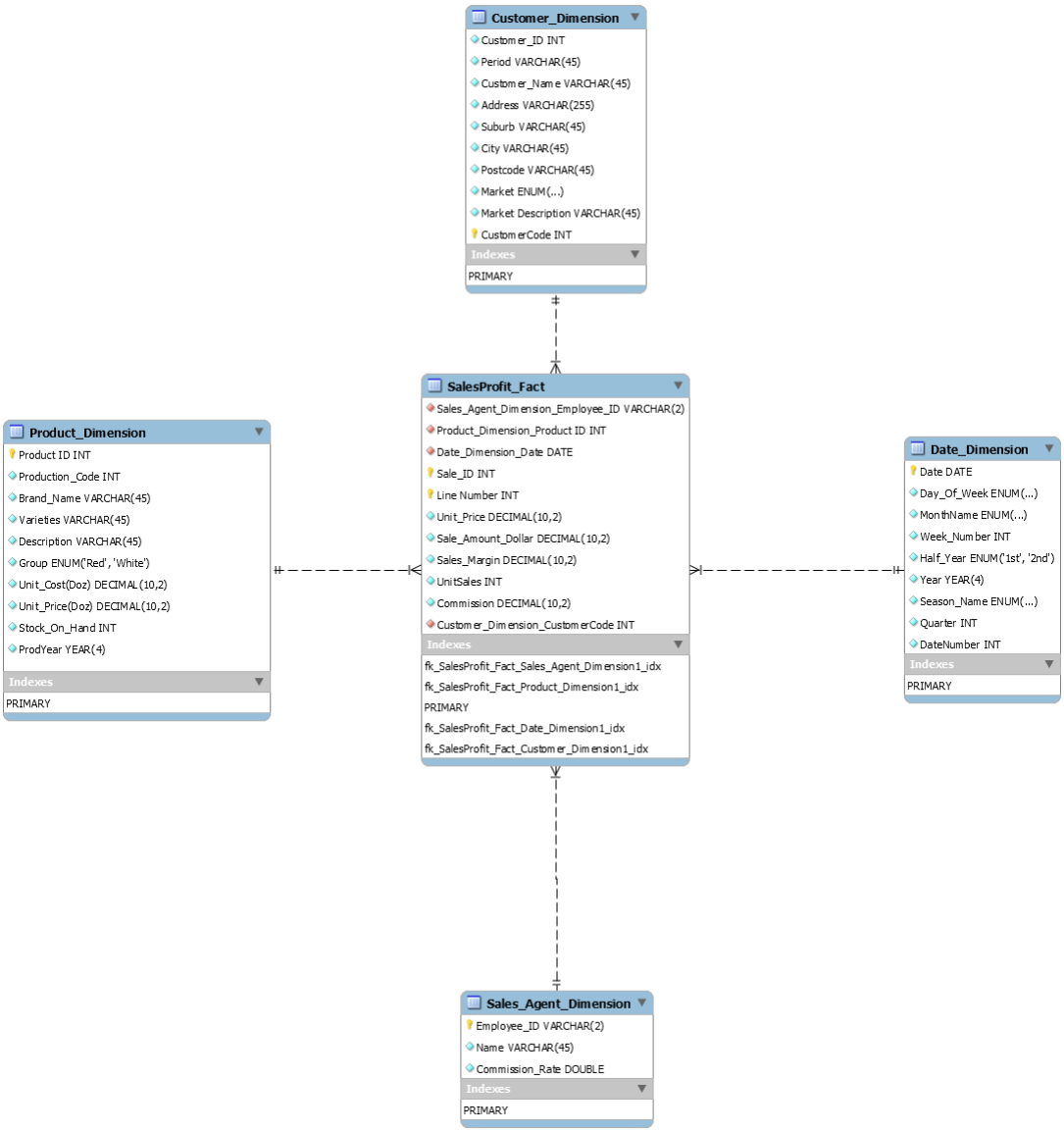
Sales_Agent_I	Product_Dime ID	Date_Dimension_Date	Sale_ID	Line Numbe	Unit_Price	Sale_Amount_Dollar	Sales_Margin	UnitSales	Commission	Customer_Dimension_CustomerCode
M2	21	2017-06-29	365	1	94.00	10998.00	1989.00	117	329.94	38
D2	14	2017-07-02	366	1	90.00	3150.00	595.00	35	252.00	32
B3	8	2017-07-02	367	1	133.00	3591.00	594.00	27	251.37	37
D2	21	2017-07-03	368	1	94.00	9870.00	1785.00	105	789.60	29
D2	19	2017-07-03	369	1	104.00	6656.00	576.00	64	532.48	34
S1	21	2017-07-03	370	1	94.00	10622.00	1921.00	113	1911.96	36
D4	15	2017-07-03	371	1	137.00	7124.00	1300.00	52	284.96	37
B2	13	2017-07-03	372	1	120.00	8760.00	2774.00	73	1051.20	41
D4	13	2017-07-04	373	1	120.00	11400.00	3610.00	95	456.00	42
D3	8	2017-07-06	374	1	133.00	3591.00	594.00	27	251.37	41
B1	16	2017-07-06	375	1	104.00	9984.00	2880.00	96	1996.80	42
B2	10	2017-07-09	376	1	101.00	11514.00	2508.00	114	1381.68	27
D4	21	2017-07-09	377	1	94.00	4794.00	867.00	51	191.76	34
B2	8	2017-07-09	378	1	133.00	15428.00	2552.00	116	1851.36	37
D4	10	2017-07-09	379	1	101.00	6767.00	1474.00	67	270.68	40
D3	20	2017-07-09	380	1	115.00	4830.00	1050.00	42	338.10	42
B2	19	2017-07-10	381	1	104.00	6240.00	540.00	60	748.80	26
B2	19	2017-07-10	382	1	104.00	11128.00	963.00	107	1335.36	27
D1	11	2017-07-10	384	1	76.00	7372.00	2134.00	97	1400.68	29
B2	11	2017-07-11	386	1	76.00	3344.00	968.00	44	401.28	30
B1	12	2017-07-17	388	1	102.00	9384.00	1748.00	92	1876.80	28
B1	18	2017-07-17	389	1	74.00	7178.00	1649.00	97	1435.60	30
B1	8	2017-07-17	390	1	133.00	13566.00	2244.00	102	2713.20	33
B1	8	2017-07-17	391	1	133.00	11438.00	1892.00	86	2287.60	37
D4	17	2017-07-17	392	1	114.00	10260.00	2520.00	90	410.40	42
B2	18	2017-07-18	393	1	74.00	2812.00	646.00	38	337.44	24
B2	11	2017-07-18	394	1	76.00	3192.00	924.00	42	383.04	25
B1	11	2017-07-18	395	1	76.00	8436.00	2442.00	111	1687.20	26
D1	17	2017-07-18	396	1	114.00	8322.00	2044.00	73	1581.18	32
M2	17	2017-07-18	397	1	114.00	9006.00	2212.00	79	270.18	42
B2	14	2017-07-19	398	1	90.00	2340.00	442.00	26	280.80	27
B1	13	2017-07-19	399	1	120.00	8880.00	2812.00	74	1776.00	29
M1	21	2017-07-19	400	1	94.00	9964.00	1802.00	106	896.76	37
B1	21	2017-07-26	401	1	94.00	8930.00	1615.00	95	1786.00	29
S2	16	2017-07-26	402	1	104.00	5408.00	1560.00	52	270.40	37
B2	19	2017-07-26	403	1	104.00	11024.00	954.00	106	1322.88	42
S2	20	2017-07-27	405	1	115.00	8050.00	1750.00	70	402.50	38
D2	20	2017-07-30	406	1	115.00	4715.00	1025.00	41	377.20	29
B3	20	2017-07-30	407	1	115.00	12190.00	2650.00	106	853.30	39
D2	10	2017-07-31	408	1	101.00	4141.00	902.00	41	331.28	35
S1	17	2017-08-02	410	1	114.00	12768.00	3136.00	112	2298.24	24
S1	21	2017-08-02	411	1	94.00	5734.00	1037.00	61	1032.12	38
B1	9	2017-08-02	413	1	100.00	9100.00	3003.00	91	1820.00	39

3. Redesign of the data warehouse

Previous Design:



Current Design:



Basically, the overall structure is similar.

For the date dimension table, it replaces date_id with date as primary key, as the date as natural key is better to be the primary key for date_dimension considering the uniqueness.

For sales_agent_dimension, it will delect the first name and last name fields and add full name field to replace them. It will save more storage.

For customer_dimension, as the address and its own derived values as well as market is managed in SCDs type two. It is different from the customer name. It is going to implement new suggorate key – CustomerCode to replace customer id as primary key. With considering status of customer data, it also put new column – period to identify the time for the updated customer data. However, for the further update on the customer dimension, it probably requires to add new field for version, date start from and end date to identify the valid period of customer data.

For SalesProfit_Fact, it is going to replace the surrogate key – sales id with sales id and line number as the primary key. It also deletes cost of good sales and add unit price to better observe the change of price.

4. Data Dictionary

Salesprofit_fact:

Attribute	Description	Source	Key
Sales ID	Unique identifier for each sale in salesprofit_fact	Salesitem.xlsx from sales system	Primary Key
Line Number	Unique identifier for each line in sale in salesprofit_fact	Salesitem.xlsx from sales system	Primary Key
Unit Price	The price of each product	Product.xlsx form sales system	
CustomerCode	Unique identifier for customer dimension	Surrogate key (Period and Customer ID)	Foreign Key
Date	Unique identifier for date dimension	Date.xlsx from external sources	Foreign Key

There are several derived values which are total sales of each line as sale_amount_dollar and sales margin as well as the commission of each sales line. To put it here before querying, it does save much memory to join tables when querying.

Customer Dimension:

Attribute	Description	Source	Key
Period	Identifier for each customer data	Customer File Name	
Customer ID	Unique identifier for customer	Customer.xlsx from sales system	
CustomerCode	Unique identifier for customer dimension	Surrogate key (Period and Customer ID)	Primary Key

Further alteration:

Attribute	Description	Source	Key
Version	Version identifier for the table	ETL	
DateFrom	Valid Start Date for the table	ETL	
DateEnd	Valid End Date for the table	ETL	

Critically, the customer dimension can be input more easily with batch inputs through Excel

Inputs. It does only consider the sequence of input data without considering the update of customer data and use the updated customer data as reference for analysis. However, from my point of view, if customers change his market later from domestic to international or from Melbourne to London, all of history sales in domestic will be regarded as international sales. It does not reflect real profitable market in the analysis. Therefore, it is necessary to consider dimension changes in the ETL.

Date Dimension

Attribute	Description	Source	Key
Date	Unique identifier for date dimension	Customer File Name	Primary Key

Appendix 1: Work Breakdown

Hangyu Pan:

Use Pentaho to implement design of ETL; Redesign Data Warehouse based on the data provided; Describe additional data dictionary