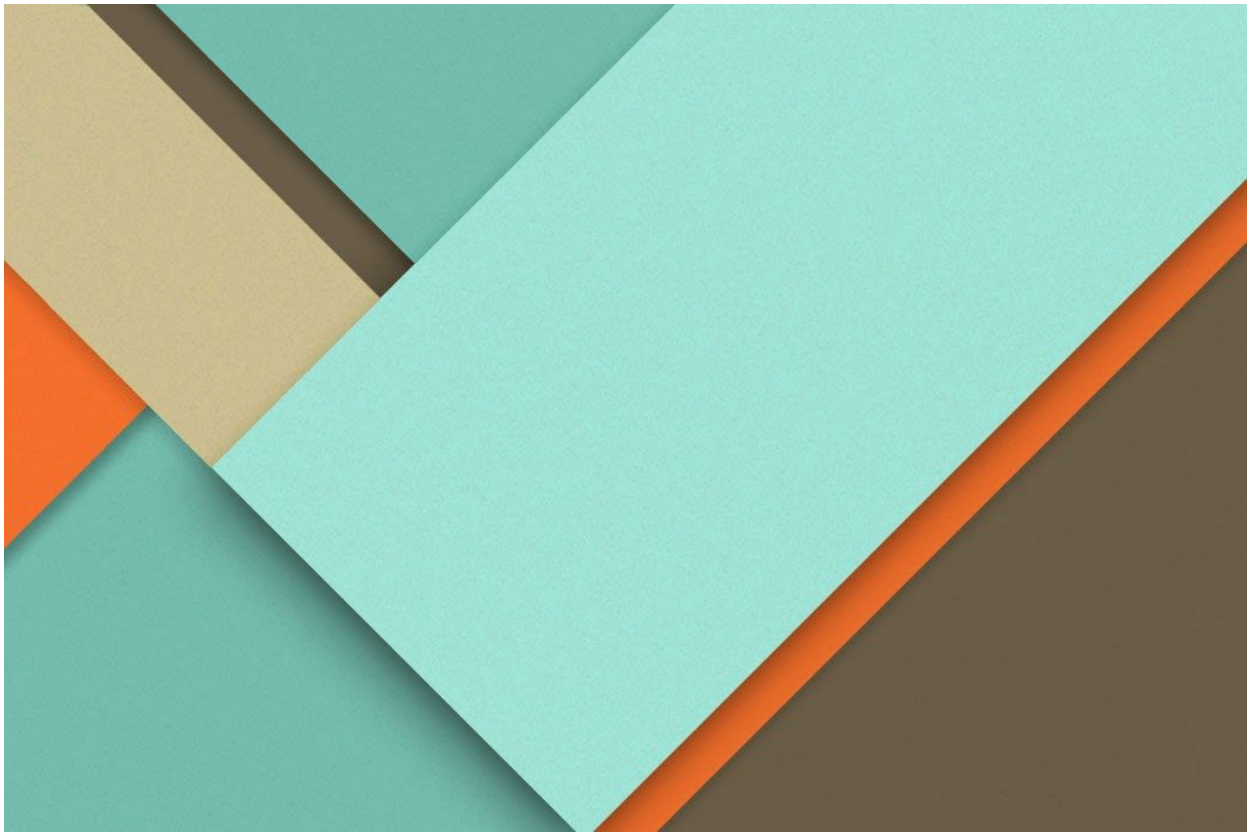


Projet Data Integration

16.01.2021

Réalisé par: BELLARABI Ibtihal

Vue d'ensemble

Parmi les étapes les plus importantes du business Intelligence est la collecte et l'alimentation des données en utilisant les outils ETL (EXTRACTION- TRANSFORMATION-lOAD).

Il consiste à collecter les infos nécessaires depuis les différentes sources extraites, transformées et chargées dans un entrepôt de données.

Objectifs

1. Se familiariser avec les tâches d'intégration des données
2. nettoyage des données et centralisation dans une datawarehouse

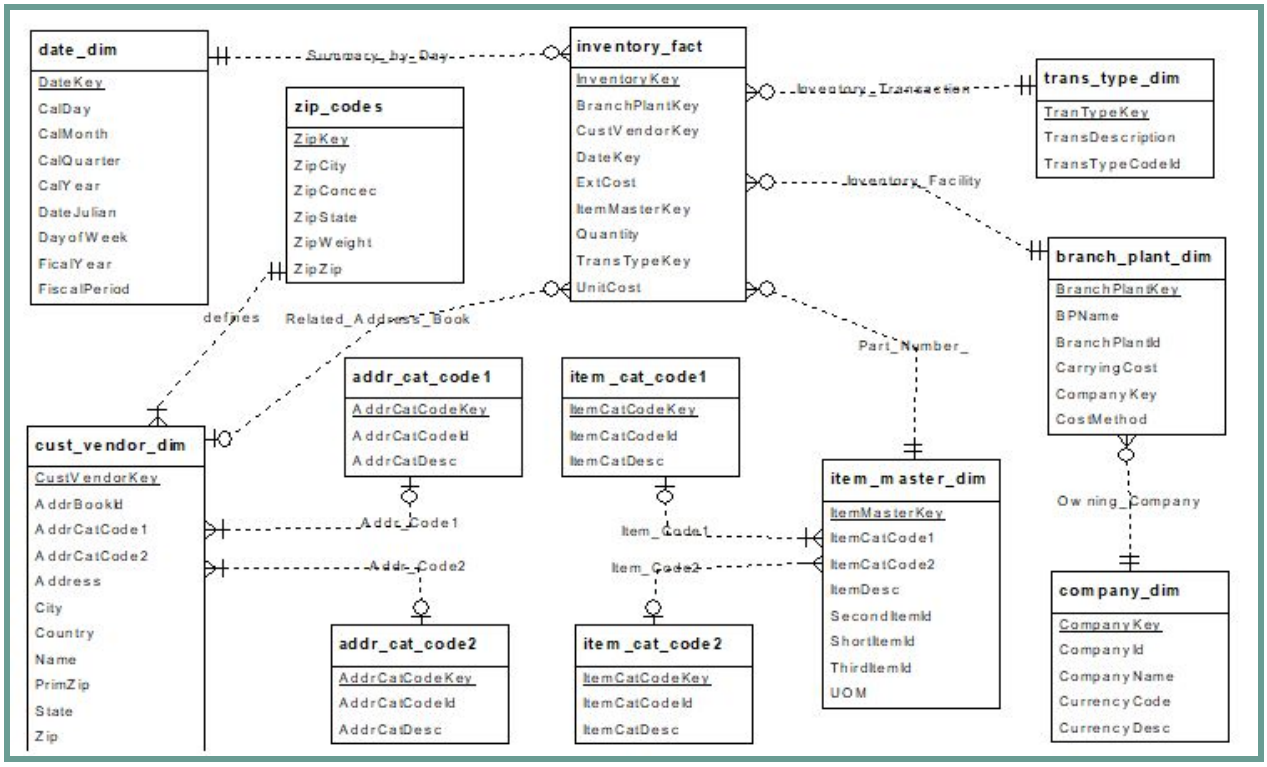
Outils

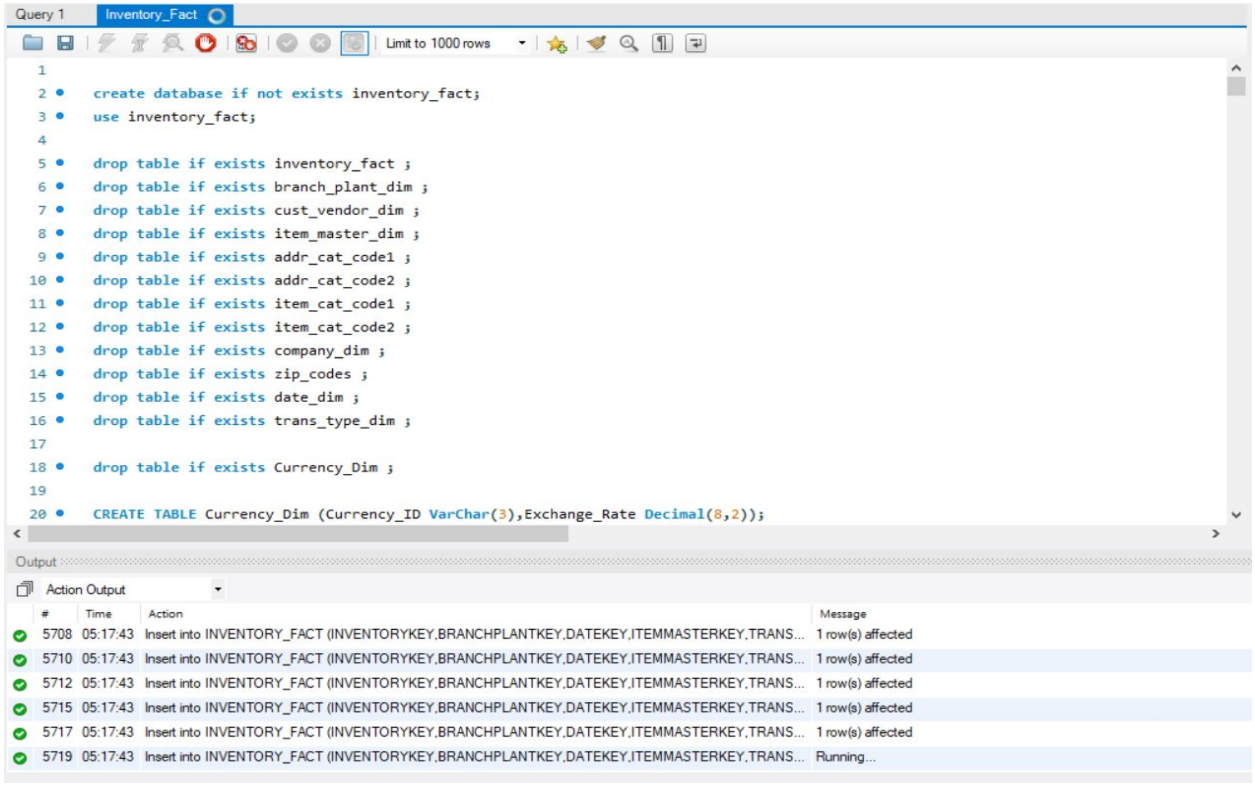
Pentaho Data Integration (interface graphique Spoon)

Grandes étapes

I. Création de la data warehouse (de la table Inventory_Fact)

Snowflake Schema



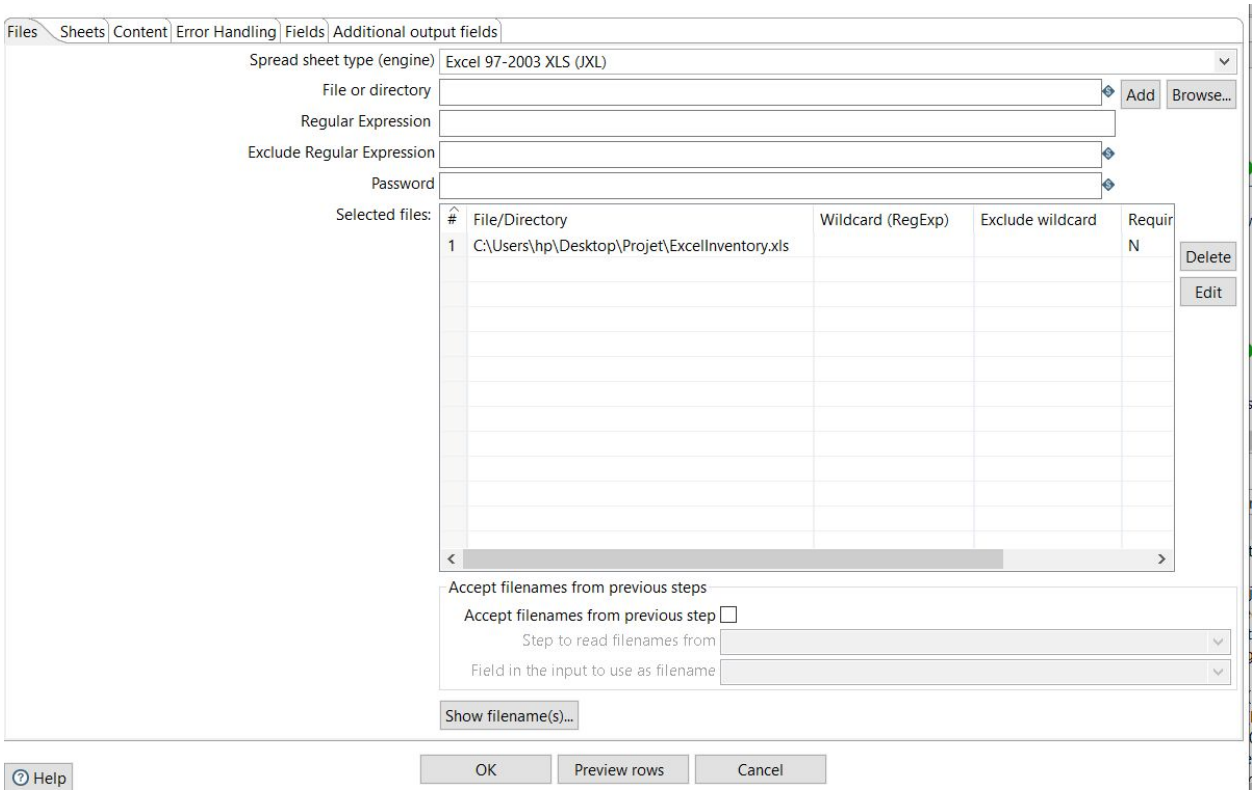

Query 1
Inventory_Fact
Limit to 1000 rows
create database if not exists inventory_fact;
use inventory_fact;
drop table if exists inventory_fact ;
drop table if exists branch_plant_dim ;
drop table if exists cust_vendor_dim ;
drop table if exists item_master_dim ;
drop table if exists addr_cat_code1 ;
drop table if exists addr_cat_code2 ;
drop table if exists item_cat_code1 ;
drop table if exists item_cat_code2 ;
drop table if exists company_dim ;
drop table if exists zip_codes ;
drop table if exists date_dim ;
drop table if exists trans_type_dim ;
drop table if exists Currency_Dim ;
CREATE TABLE Currency_Dim (Currency_ID VarChar(3),Exchange_Rate Decimal(8,2));
Output
Action Output
Time Action Message
5708 05:17:43 Insert into INVENTORY_FACT (INVENTORYKEY,BRANCHPLANTKEY,DATEKEY,ITEMMASTERKEY,TRANS... 1 row(s) affected
5710 05:17:43 Insert into INVENTORY_FACT (INVENTORYKEY,BRANCHPLANTKEY,DATEKEY,ITEMMASTERKEY,TRANS... 1 row(s) affected
5712 05:17:43 Insert into INVENTORY_FACT (INVENTORYKEY,BRANCHPLANTKEY,DATEKEY,ITEMMASTERKEY,TRANS... 1 row(s) affected
5715 05:17:43 Insert into INVENTORY_FACT (INVENTORYKEY,BRANCHPLANTKEY,DATEKEY,ITEMMASTERKEY,TRANS... 1 row(s) affected
5717 05:17:43 Insert into INVENTORY_FACT (INVENTORYKEY,BRANCHPLANTKEY,DATEKEY,ITEMMASTERKEY,TRANS... 1 row(s) affected
5719 05:17:43 Insert into INVENTORY_FACT (INVENTORYKEY,BRANCHPLANTKEY,DATEKEY,ITEMMASTERKEY,TRANS... Running...

II. ETL Opérations

1. Data Source

we need to perform cleaning operations on two data sources. Both data sources provide facts for the *Inventory_Fact* table of the inventory data warehouse. We need to create a transform for each data source.

we create the 



Microsoft Excel input

Step name: Inventory_EXCEL

Files | Sheets | Content | Error Handling | Fields | Additional output fields

#	Name	Type	Length	Precision	Trim type	Repeat	Format	Currency
1	BranchPlantKey	Integer	-1	-1	none	N		
2	Date	Date	-1	-1	none	N		
3	ItemMasterKey	Integer	-1	-1	none	N		
4	TransTypeKey	Integer	-1	-1	none	N		
5	CustVendorKey	Integer	-1	-1	none	N		
6	UnitCost	Number	-1	-1	none	N		
7	Currency	String	-1	-1	none	N		
8	Quantity	Integer	-1	-1	none	N		
9								
1..								
1..								

Get fields from header row...

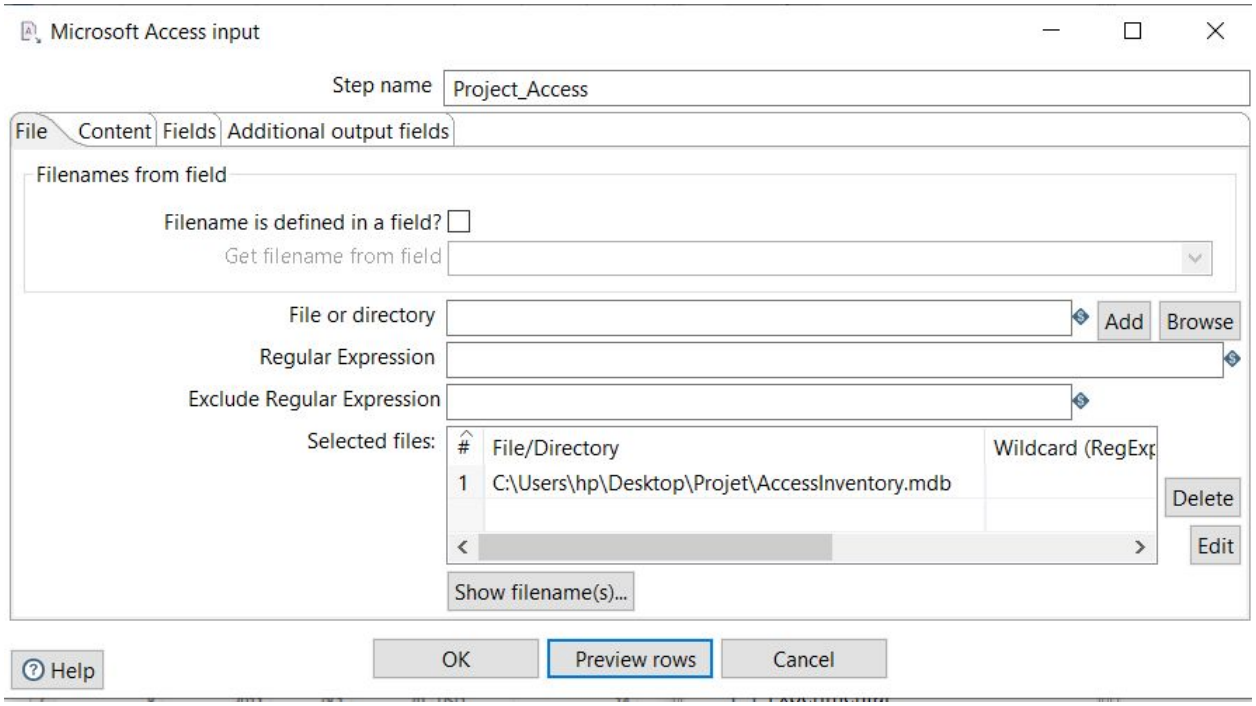
Help | OK | Preview rows | Cancel

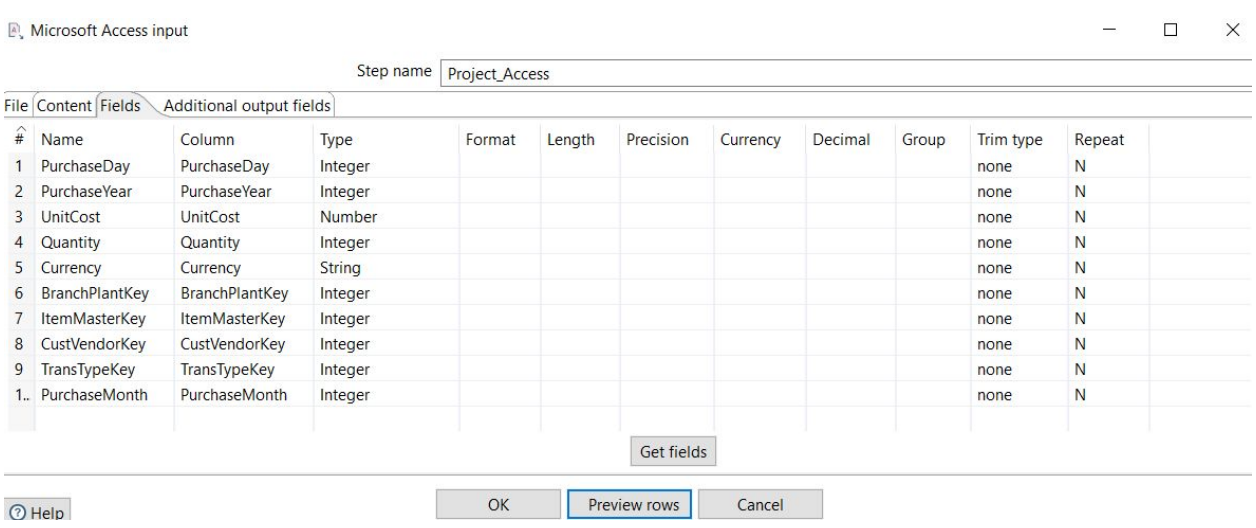
Examine preview data

Rows of step: Inventory_EXCEL (15 rows)

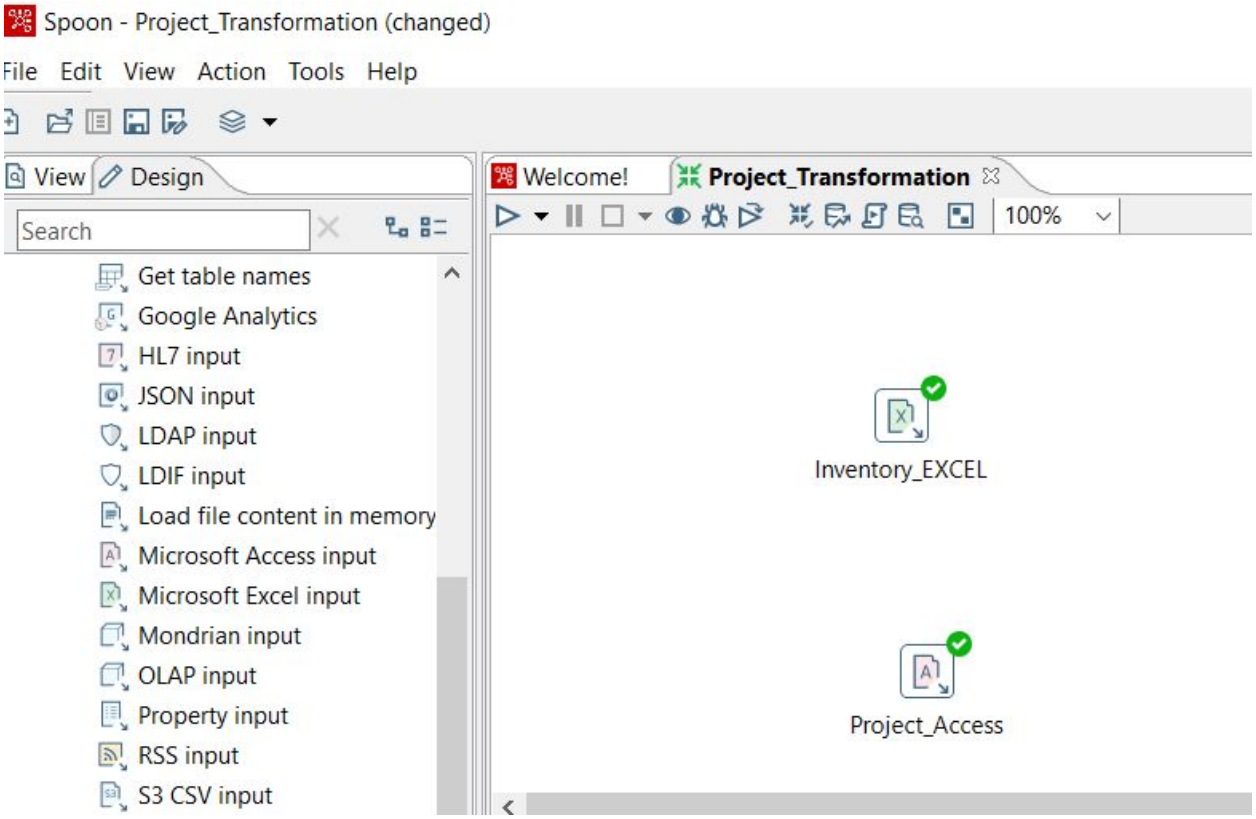
#	BranchPlantKey	Date	ItemMasterKey	TransTypeKey	CustVendorKey	UnitCost	Currency	Quantity
1	1	2012/08/06 00:00:00.000	19	1	5	11,22	USD	10
2	3	2012/08/13 00:00:00.000	17	3	7	22,33	JPY	20
3	5	2012/08/20 00:00:00.000	9	5	1	33,11	USD	30
4	7	2012/08/27 00:00:00.000	7	1	3	33,22	JPY	40
5	9	2012/09/03 00:00:00.000	15	1	11	22,01	GBP	50
6	11	2012/09/10 00:00:00.000	5	3	15	11,03	USD	20
7	13	2013/09/18 00:00:00.000	9	5	13	15,04	GBP	22
8	15	2012/09/24 00:00:00.000	<null>	3	7	17,09	USD	18
9	17	2012/10/01 00:00:00.000	5	3	5	33,02	JPY	15
1..	19	2013/10/08 00:00:00.000	7	1	1	22,11	USD	16
1..	5	2012/10/15 00:00:00.000	1	3	15	17,11	JPY	17
1..	7	2012/10/22 00:00:00.000	3	1	13	21,22	USD	15
1..	9	2012/10/29 00:00:00.000	5	5	11	33,21	USD	18
1..	13	2012/11/05 00:00:00.000	15	<null>	9	44,11	JPY	13
1..	15	2012/11/12 00:00:00.000	13	1	99	12,11	USD	14

we create the 

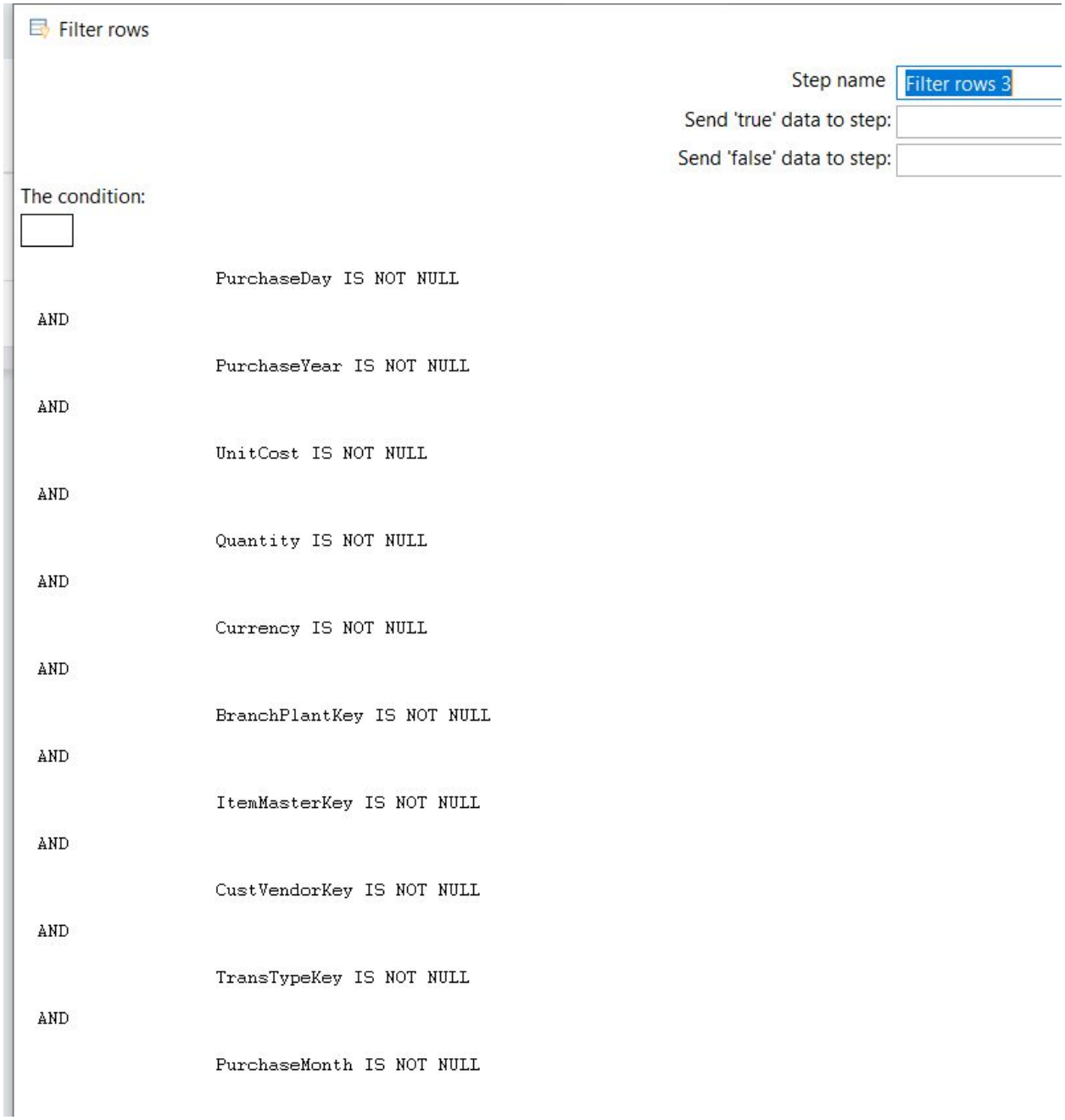


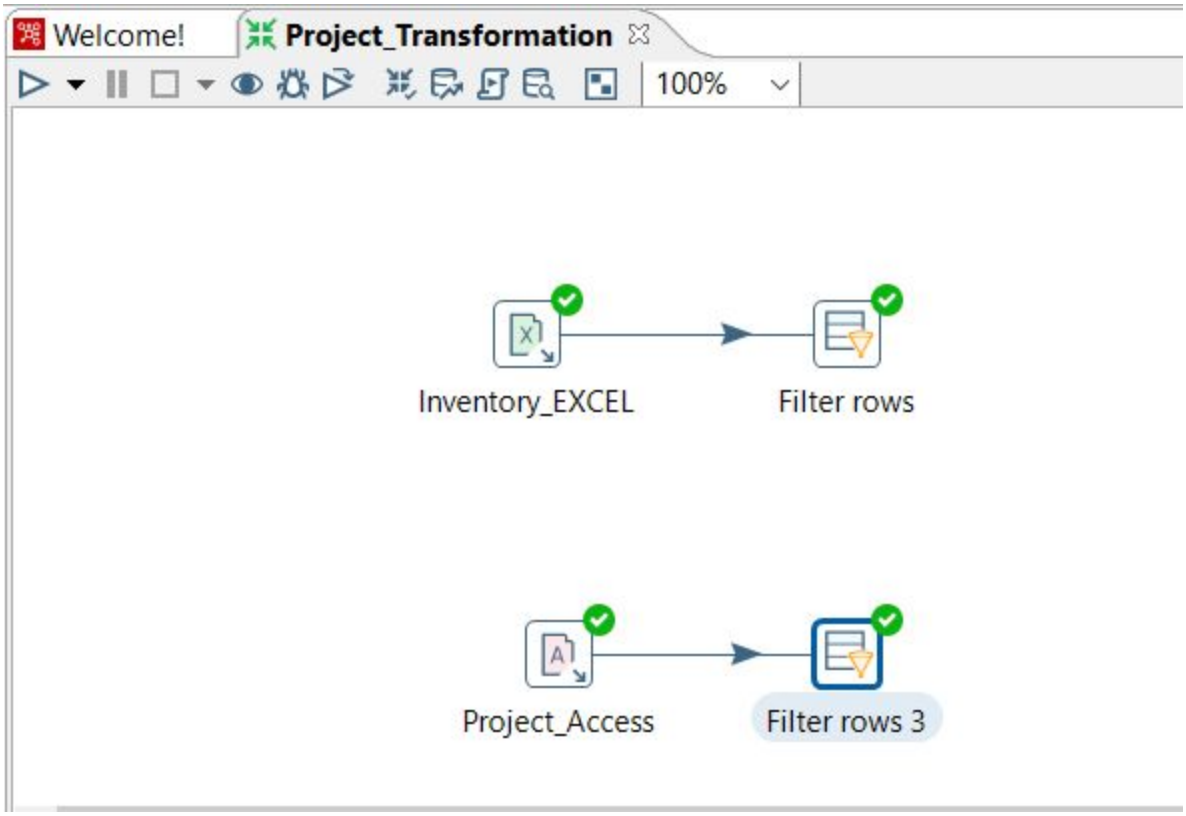


#	PurchaseDay	PurchaseYear	UnitCost	Quantity	Currency	BranchPlantKey	ItemMasterKey	CustVendorKey	TransTypeKey	PurchaseMonth
1	10	2011	45,2	10	USD	2	2	4	2	1
2	28	2013	16,1	20	JPY	4	8	18	4	2
3	14	2011	15,7	30	GBP	6	6	14	4	2
4	21	2011	25,6	40	EUR	8	<null>	16	4	3
5	4	2011	54,5	50	EUR	10	14	18	4	4
6	3	2011	24,3	60	USD	12	1	12	2	1
7	8	2011	78,1	70	USD	14	8	14	4	2
8	31	2011	14,45	80	USD	16	4	6	2	1
9	18	2011	22,0	90	USD	18	20	8	2	4
1..	11	2011	12,56	12	JPY	20	18	<null>	2	4
1..	28	2011	18,26	25	JPY	18	12	2	2	3
1..	7	2011	19,11	35	USD	16	10	99	2	3
1..	14	2011	20,77	45	EUR	16	10	12	2	3
1..	17	2011	26,89	55	JPY	14	2	10	4	1
1..	21	2011	25,36	65	USD	8	6	10	4	2
1..	7	2011	14,12	75	GBP	2	4	8	4	2

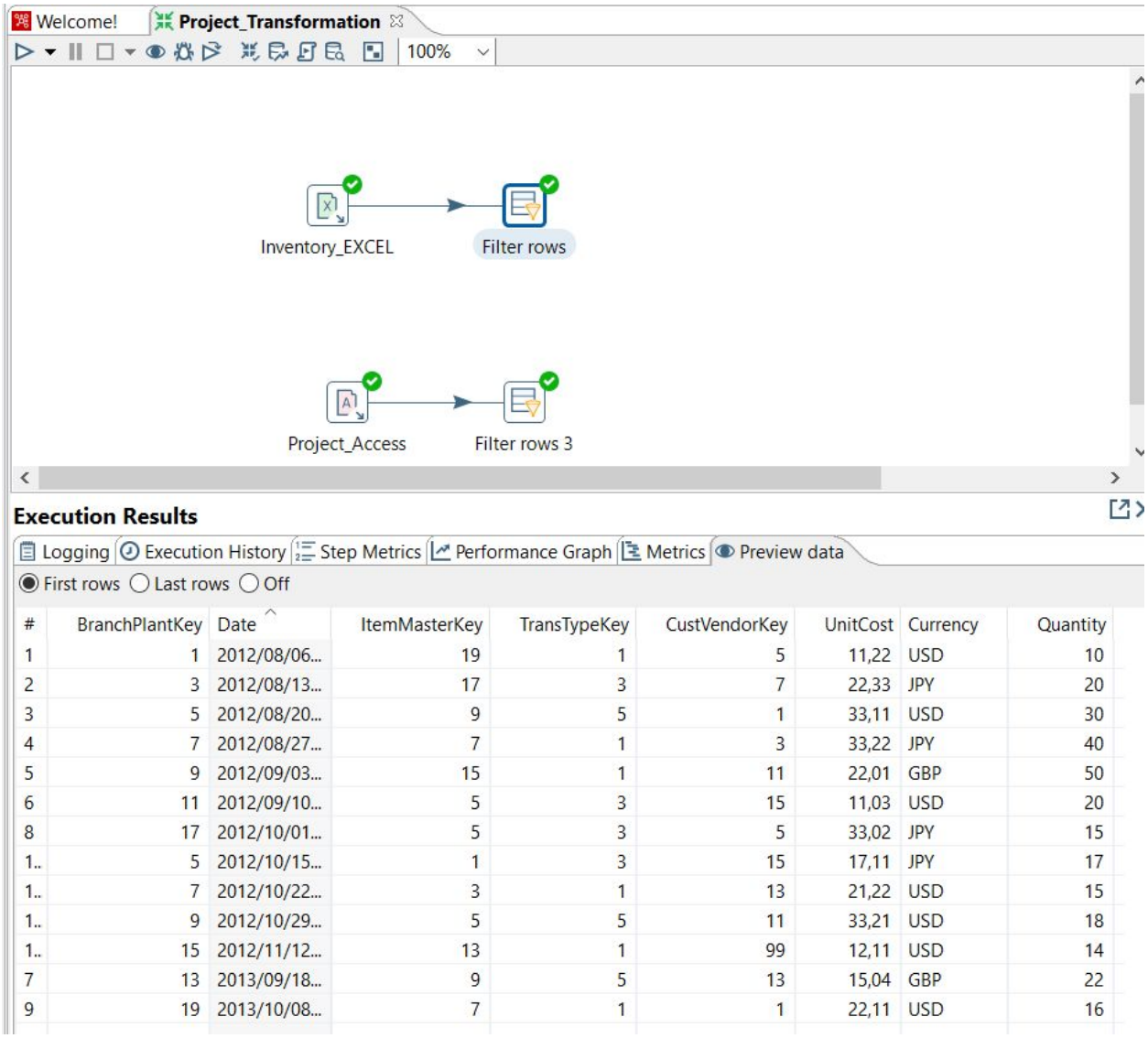


Reject a record if any field is null.




Welcome!
Project_Transformation
100%
Inventory_EXCEL
Filter rows
Project_Access
Filter rows 3

After running the transformation, this is the executing results

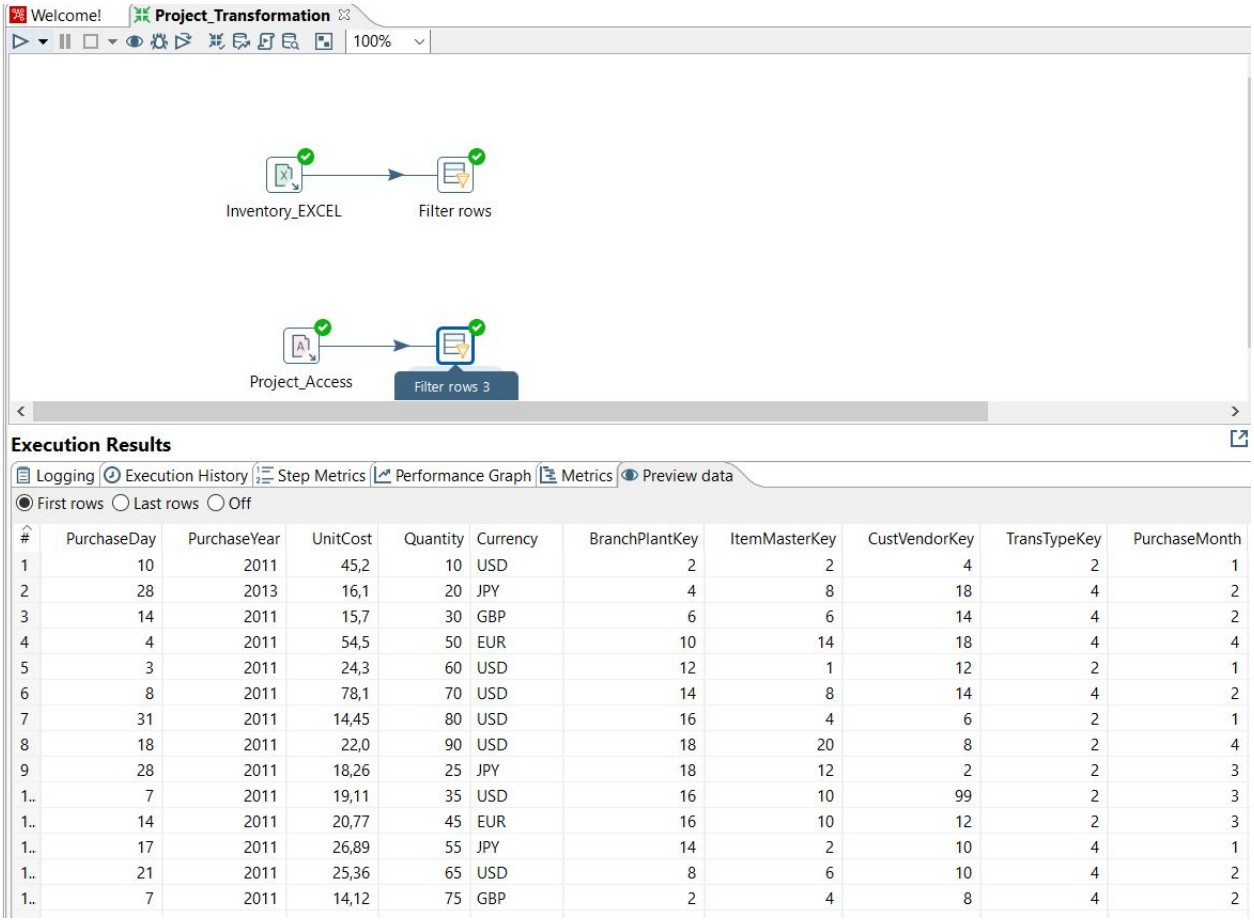


Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

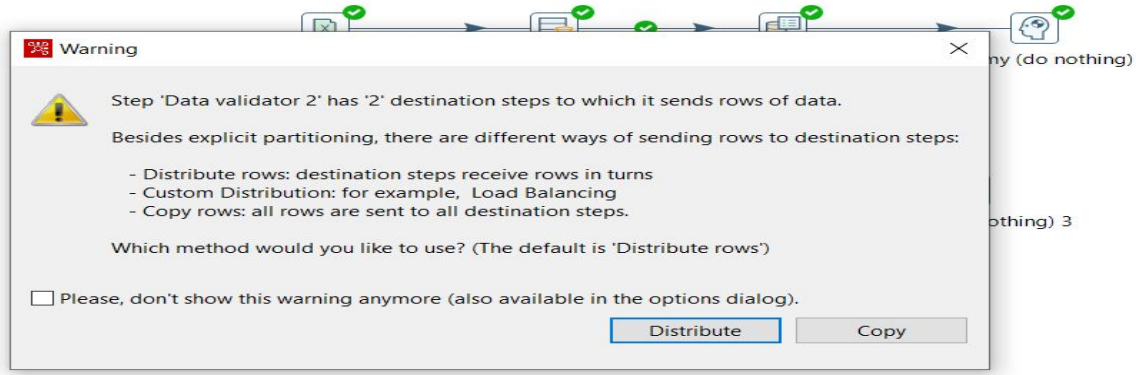
#	BranchPlantKey	Date	ItemMasterKey	TransTypeKey	CustVendorKey	UnitCost	Currency	Quantity
1	1	2012/08/06...	19	1	5	11,22	USD	10
2	3	2012/08/13...	17	3	7	22,33	JPY	20
3	5	2012/08/20...	9	5	1	33,11	USD	30
4	7	2012/08/27...	7	1	3	33,22	JPY	40
5	9	2012/09/03...	15	1	11	22,01	GBP	50
6	11	2012/09/10...	5	3	15	11,03	USD	20
8	17	2012/10/01...	5	3	5	33,02	JPY	15
1..	5	2012/10/15...	1	3	15	17,11	JPY	17
1..	7	2012/10/22...	3	1	13	21,22	USD	15
1..	9	2012/10/29...	5	5	11	33,21	USD	18
1..	15	2012/11/12...	13	1	99	12,11	USD	14
7	13	2013/09/18...	9	5	13	15,04	GBP	22
9	19	2013/10/08...	7	1	1	22,11	USD	16

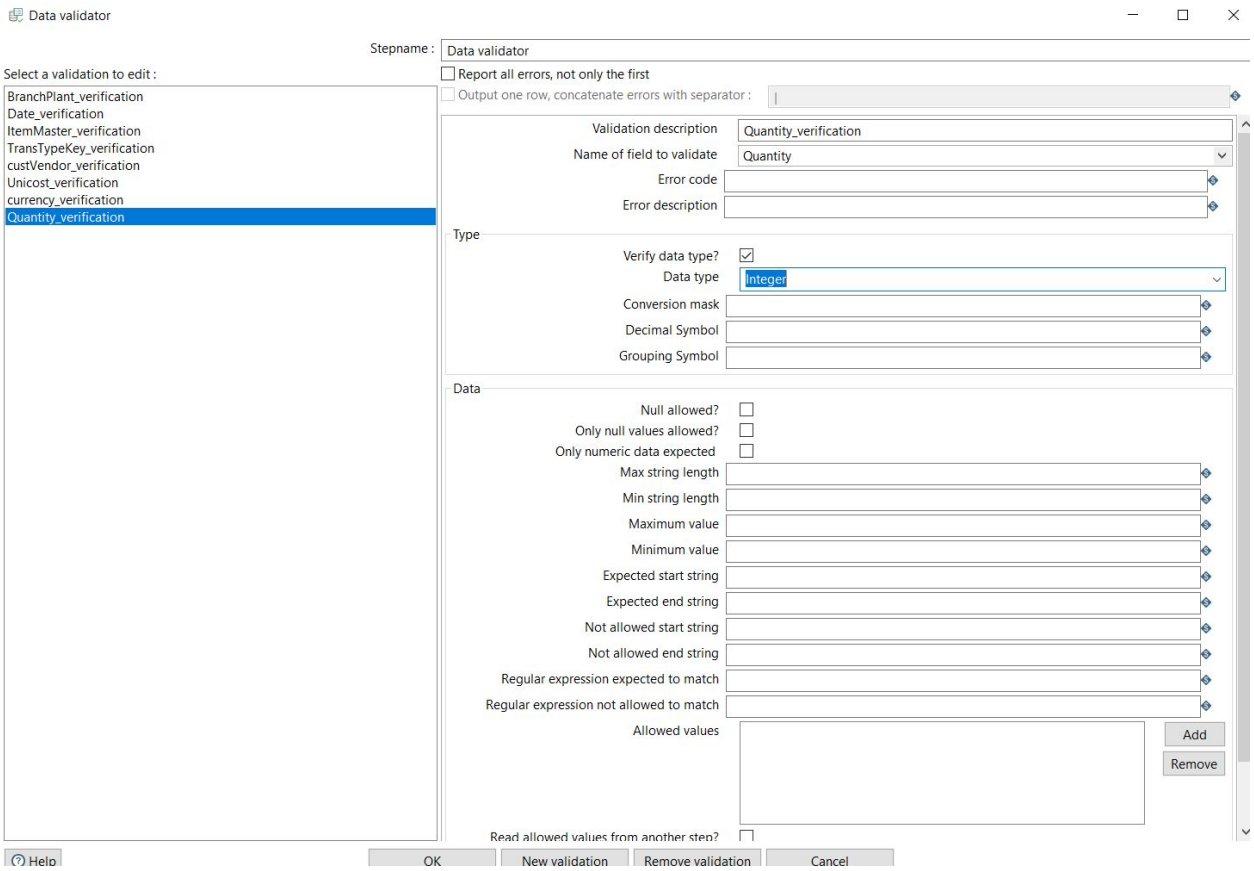


#	PurchaseDay	PurchaseYear	UnitCost	Quantity	Currency	BranchPlantKey	ItemMasterKey	CustVendorKey	TransTypeKey	PurchaseMonth
1	10	2011	45,2	10	USD	2	2	4	2	1
2	28	2013	16,1	20	JPY	4	8	18	4	2
3	14	2011	15,7	30	GBP	6	6	14	4	2
4	4	2011	54,5	50	EUR	10	14	18	4	4
5	3	2011	24,3	60	USD	12	1	12	2	1
6	8	2011	78,1	70	USD	14	8	14	4	2
7	31	2011	14,45	80	USD	16	4	6	2	1
8	18	2011	22,0	90	USD	18	20	8	2	4
9	28	2011	18,26	25	JPY	18	12	2	2	3
1..	7	2011	19,11	35	USD	16	10	99	2	3
1..	14	2011	20,77	45	EUR	16	10	12	2	3
1..	17	2011	26,89	55	JPY	14	2	10	4	1
1..	21	2011	25,36	65	USD	8	6	10	4	2
1..	7	2011	14,12	75	GBP	2	4	8	4	2

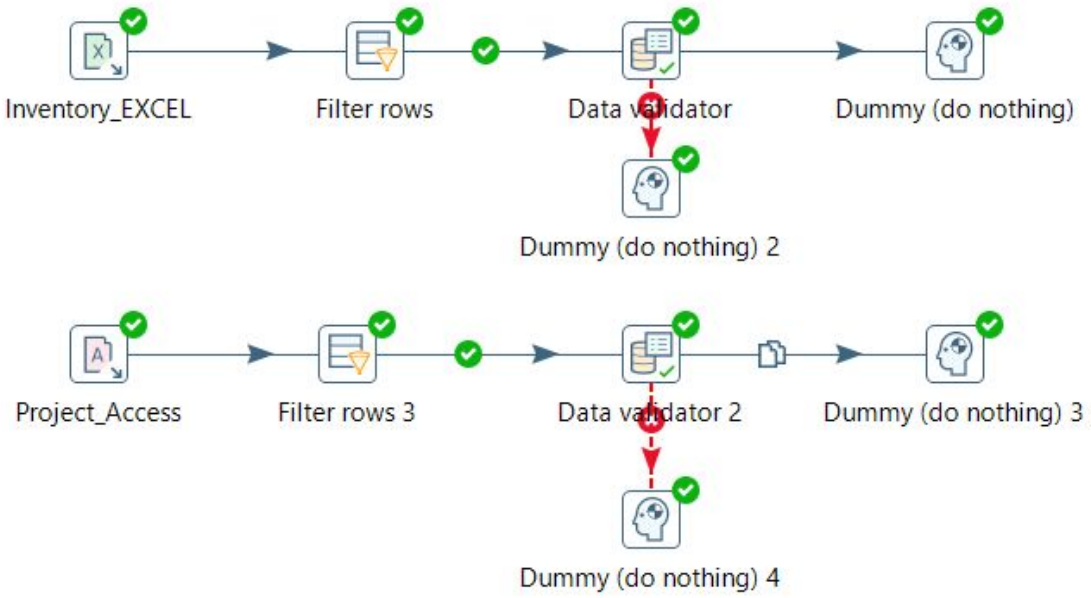
Reject a record if any field value does not match its data type.

To do the verification, we create a data validator for all the fields, it checks the type of data and filter the admitted rows

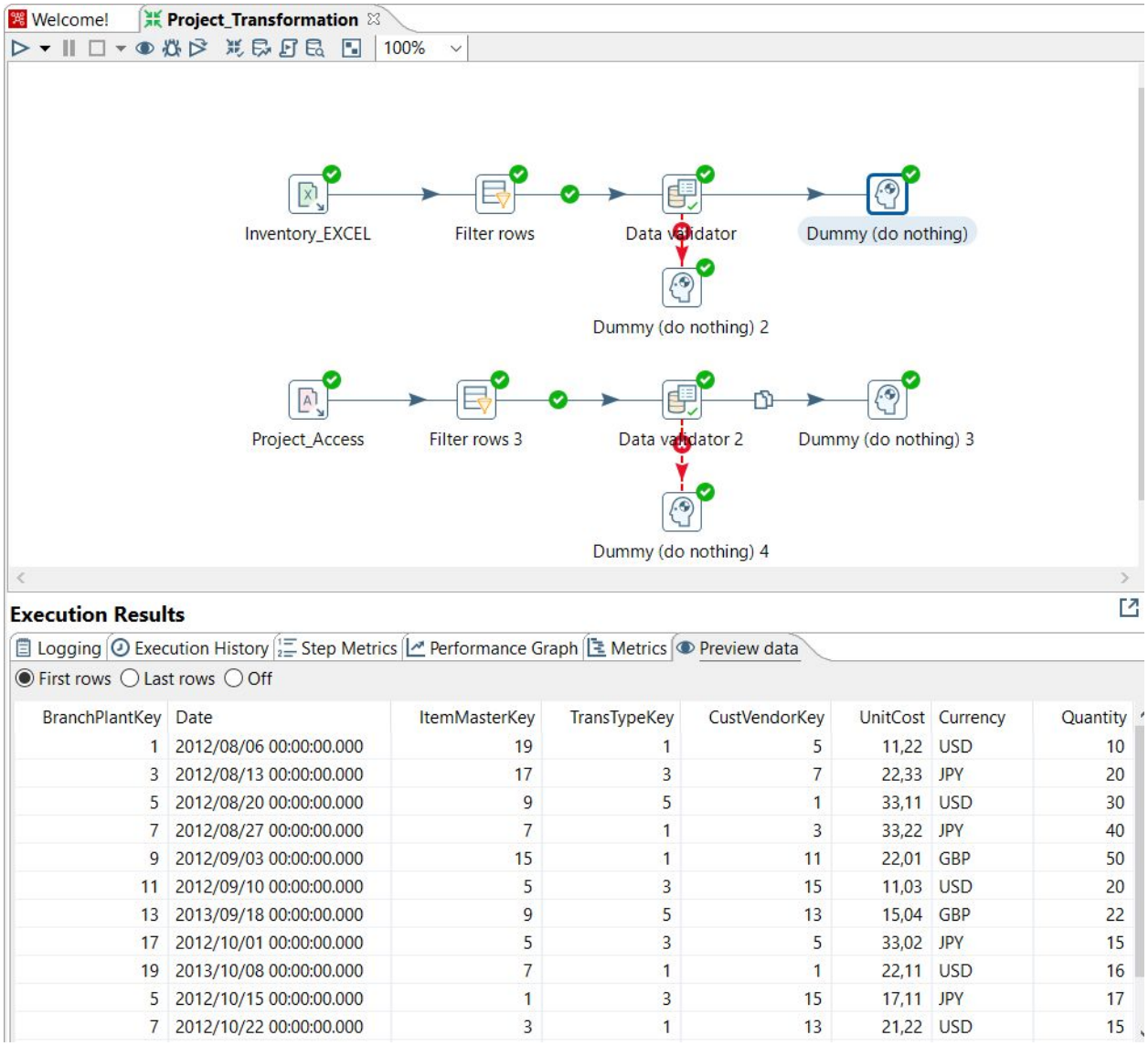




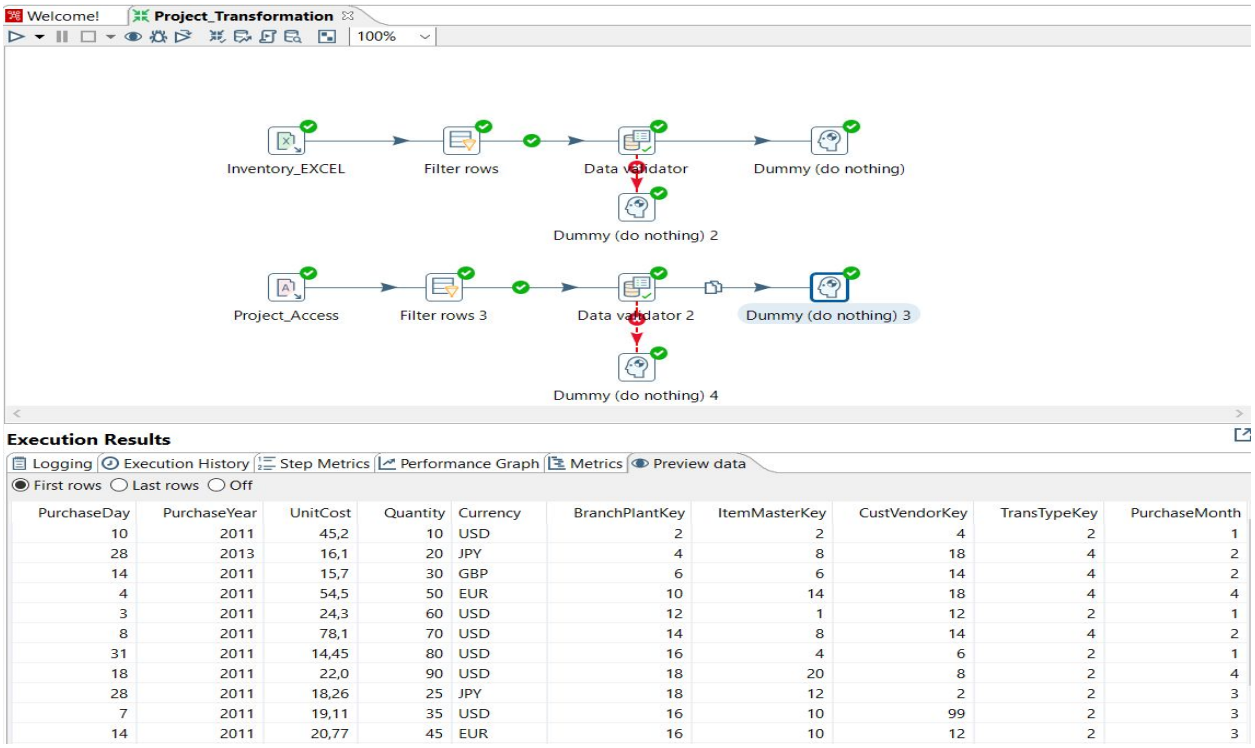
After running the transformation



All data types are suiting the type of their field



BranchPlantKey	Date	ItemMasterKey	TransTypeKey	CustVendorKey	UnitCost	Currency	Quantity
1	2012/08/06 00:00:00.000	19	1	5	11,22	USD	10
3	2012/08/13 00:00:00.000	17	3	7	22,33	JPY	20
5	2012/08/20 00:00:00.000	9	5	1	33,11	USD	30
7	2012/08/27 00:00:00.000	7	1	3	33,22	JPY	40
9	2012/09/03 00:00:00.000	15	1	11	22,01	GBP	50
11	2012/09/10 00:00:00.000	5	3	15	11,03	USD	20
13	2013/09/18 00:00:00.000	9	5	13	15,04	GBP	22
17	2012/10/01 00:00:00.000	5	3	5	33,02	JPY	15
19	2013/10/08 00:00:00.000	7	1	1	22,11	USD	16
5	2012/10/15 00:00:00.000	1	3	15	17,11	JPY	17
7	2012/10/22 00:00:00.000	3	1	13	21,22	USD	15



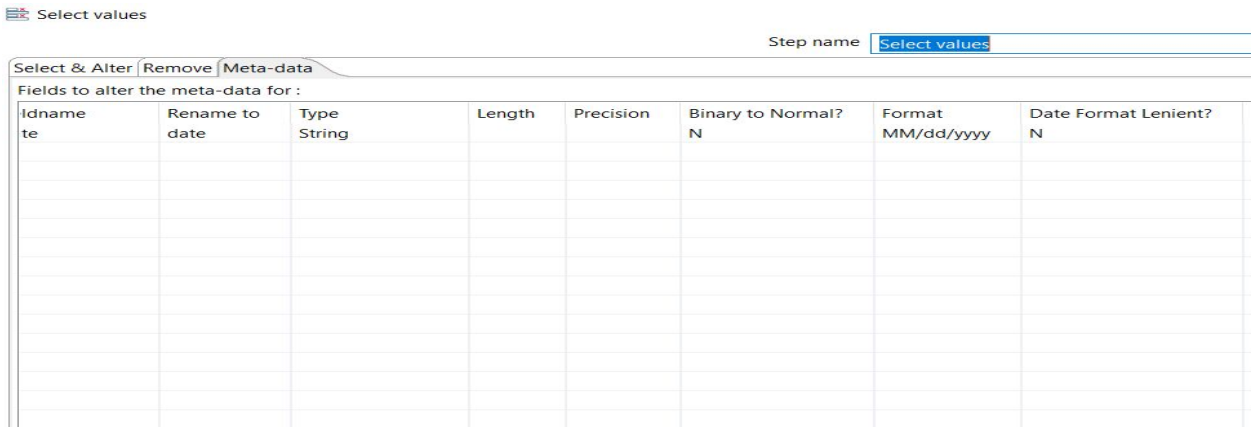
PurchaseDay	PurchaseYear	UnitCost	Quantity	Currency	BranchPlantKey	ItemMasterKey	CustVendorKey	TransTypeKey	PurchaseMonth
10	2011	45,2	10	USD	2	2	4	2	1
28	2013	16,1	20	JPY	4	8	18	4	2
14	2011	15,7	30	GBP	6	6	14	4	2
4	2011	54,5	50	EUR	10	14	18	4	4
3	2011	24,3	60	USD	12	1	12	2	1
8	2011	78,1	70	USD	14	8	14	4	2
31	2011	14,45	80	USD	16	4	6	2	1
18	2011	22,0	90	USD	18	20	8	2	4
28	2011	18,26	25	JPY	18	12	2	2	3
7	2011	19,11	35	USD	16	10	99	2	3
14	2011	20,77	45	EUR	16	10	12	2	3

Reject invalid dates: the combination of month, day, and year should be a valid date (including leap year processing) that exists in the *Date_Dim* table.

Parsing Dates with Excel Data Sources

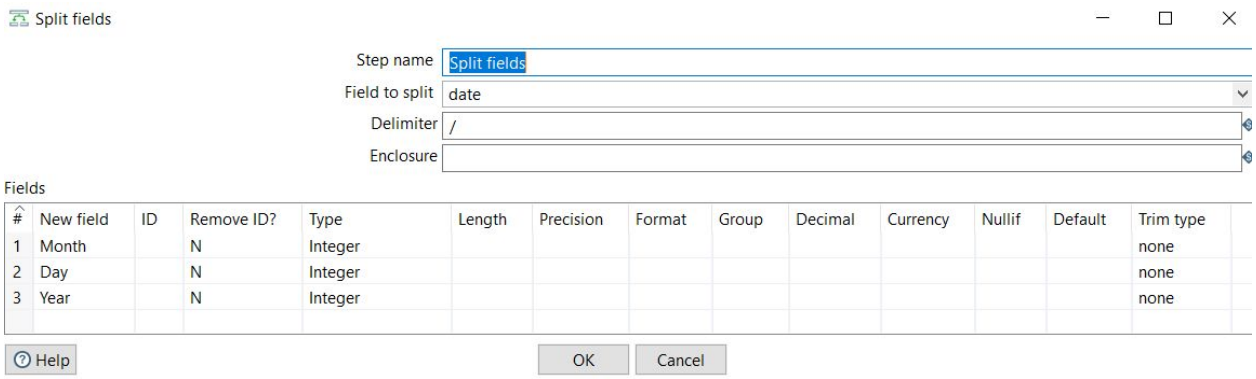
To do this task we need to parse the date field and split it into year, day, month fields , the check if they are valid or not.

we create selected values step and fill in the changes of the field

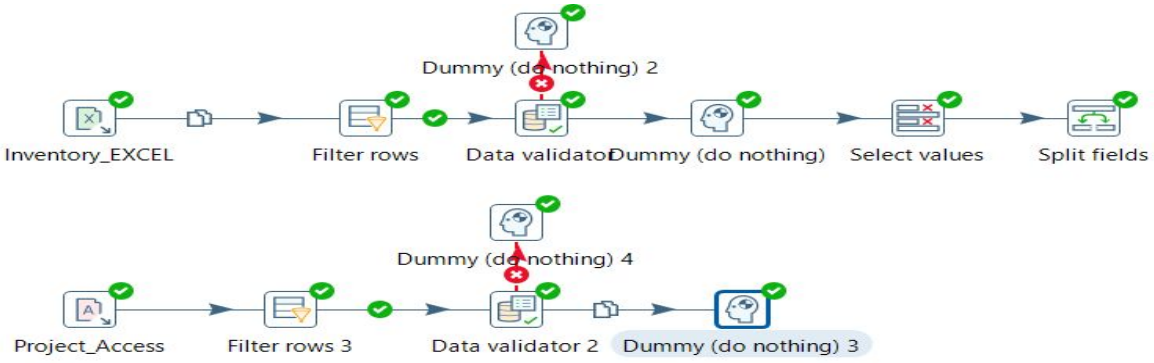


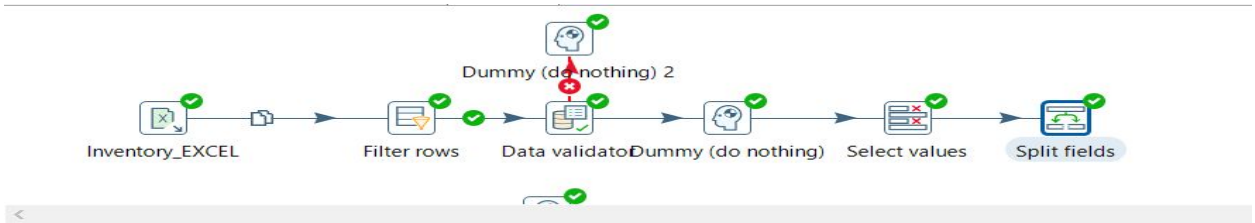
Idname te	Rename to date	Type String	Length	Precision	Binary to Normal? N	Format MM/dd/yyyy	Date Format Lenient? N

Then, we create the split step



After running the transformation





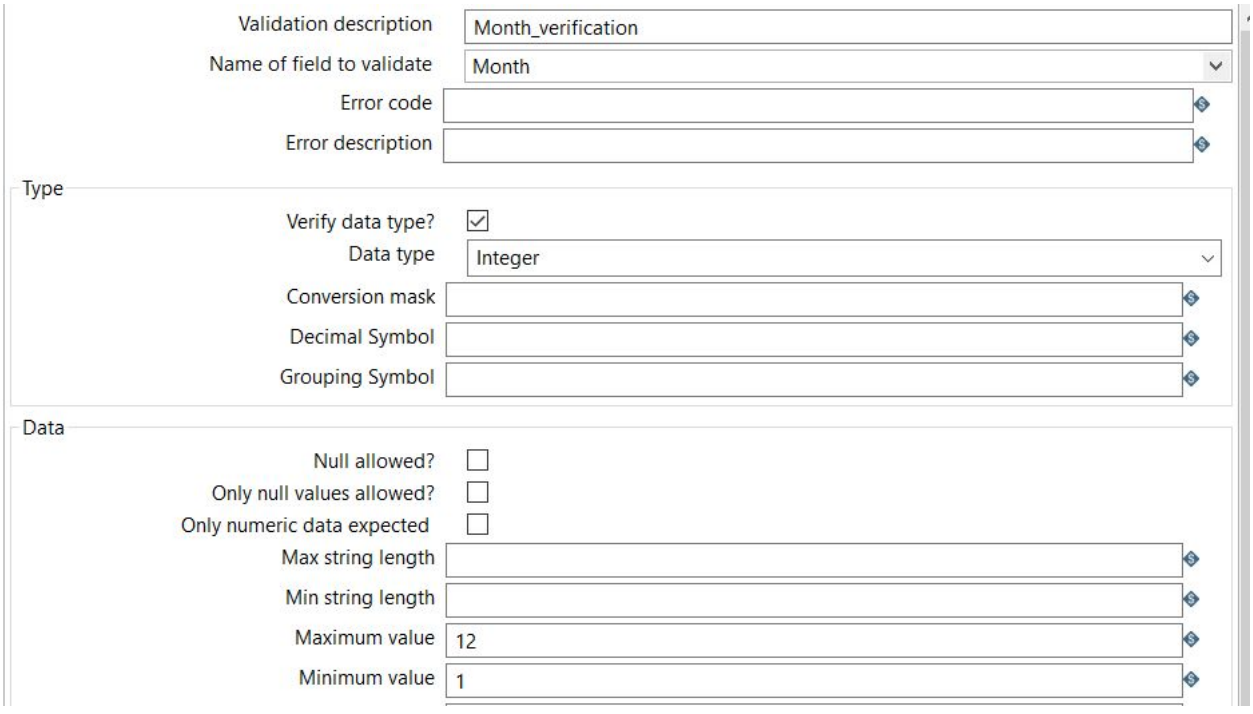
Execution Results

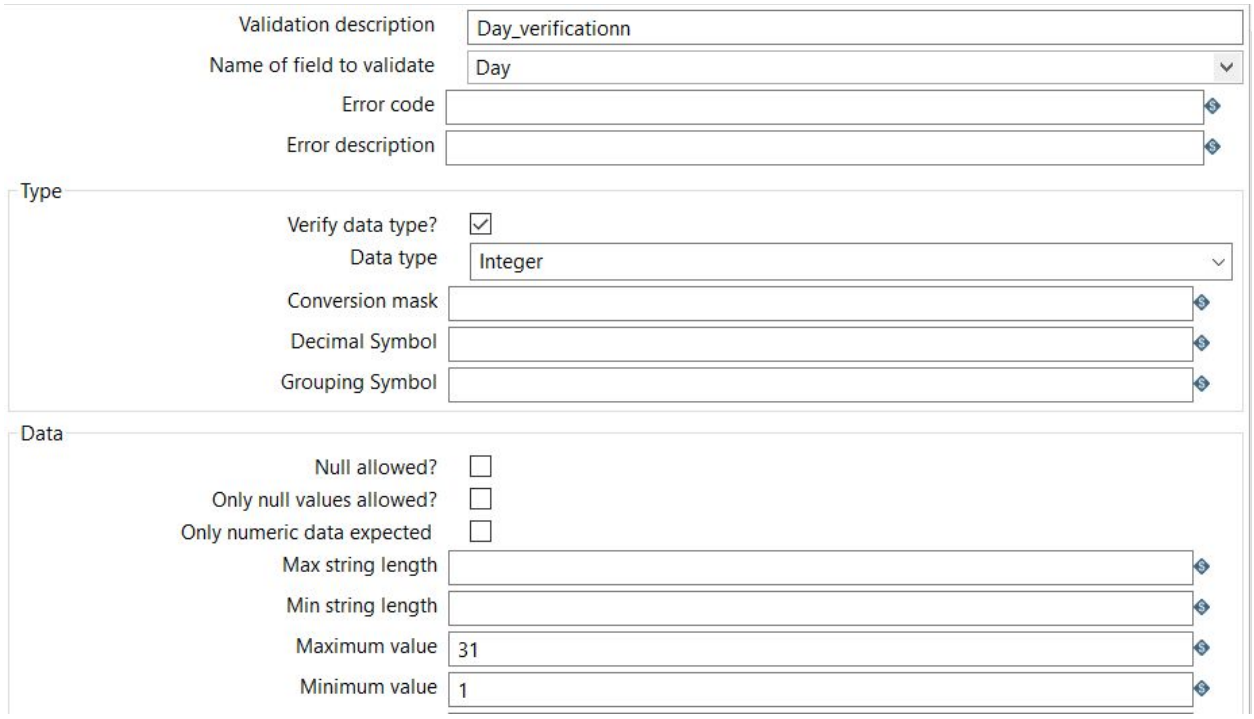
Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

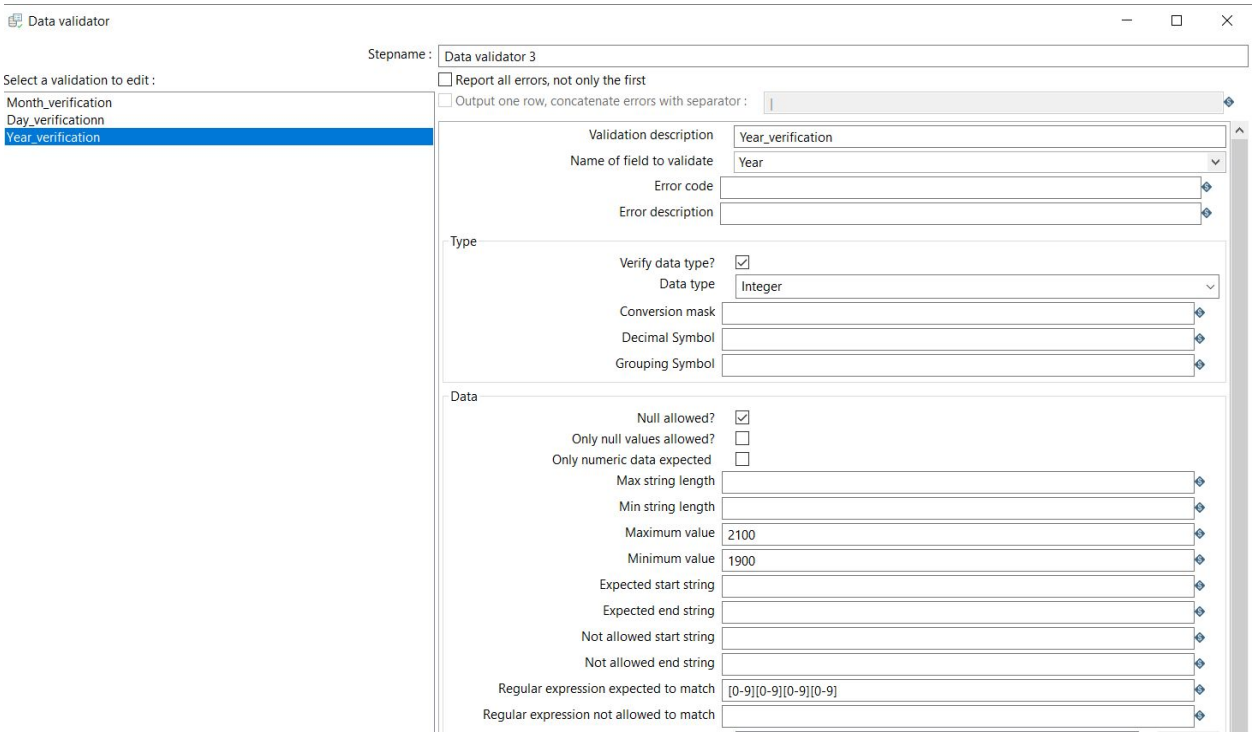
First rows | Last rows | Off

#	Month	Day	Year	BranchPlantKey	ItemMasterKey	TransTypeKey	CustVendorKey	UnitCost	Currency
1	8	6	2012	1	19	1	5	11,22	USD
2	8	13	2012	3	17	3	7	22,33	JPY
3	8	20	2012	5	9	5	1	33,11	USD
4	8	27	2012	7	7	1	3	33,22	JPY
5	9	3	2012	9	15	1	11	22,01	GBP
6	9	10	2012	11	5	3	15	11,03	USD
7	9	18	2013	13	9	5	13	15,04	GBP
8	10	1	2012	17	5	3	5	33,02	JPY
9	10	8	2013	19	7	1	1	22,11	USD
1..	10	15	2012	5	1	3	15	17,11	JPY
1..	10	22	2012	7	3	1	13	21,22	USD
1..	10	29	2012	9	5	5	11	33,21	USD
1..	11	12	2012	15	13	1	99	12,11	USD

We create a data validator to put validations conditions of the components of date

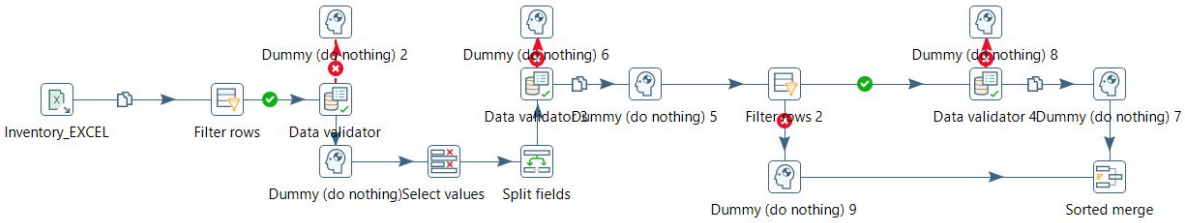


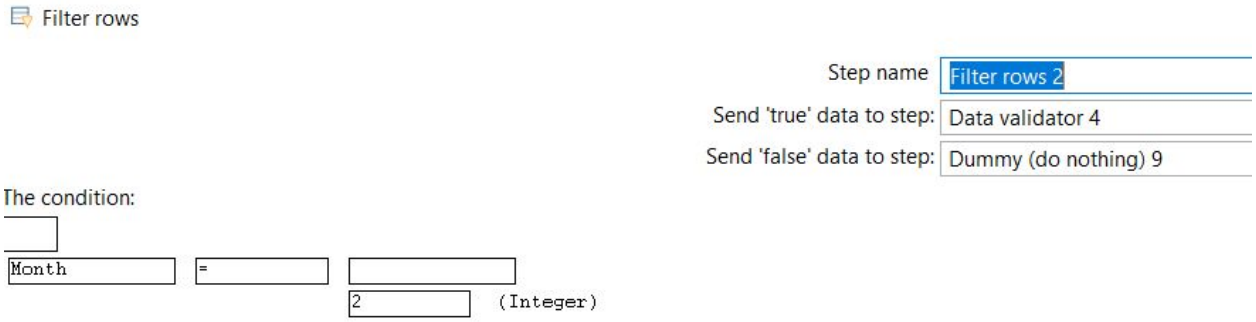

Validation description
Day_verificationn
Name of field to validate
Day
Error code
Error description
Type
Verify data type?
Data type
Integer
Conversion mask
Decimal Symbol
Grouping Symbol
Data
Null allowed?
Only null values allowed?
Only numeric data expected
Max string length
Min string length
Maximum value
31
Minimum value
1

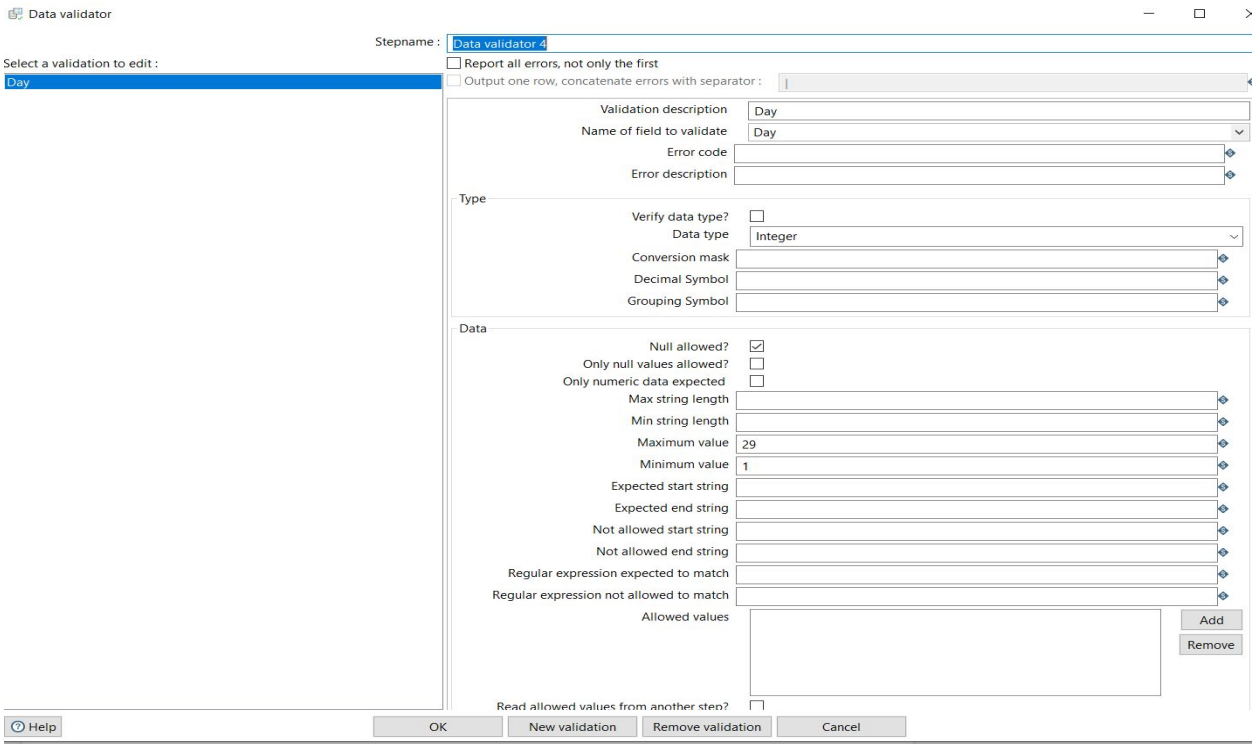

Data validator
Stepname :
Data validator 3
Select a validation to edit :
Report all errors, not only the first
Output one row, concatenate errors with separator :
Month_verification
Day_verificationn
Year_verification
Validation description
Year_verification
Name of field to validate
Year
Error code
Error description
Type
Verify data type?
Data type
Integer
Conversion mask
Decimal Symbol
Grouping Symbol
Data
Null allowed?
Only null values allowed?
Only numeric data expected
Max string length
Min string length
Maximum value
2100
Minimum value
1900
Expected start string
Expected end string
Not allowed start string
Not allowed end string
Regular expression expected to match
[0-9][0-9][0-9][0-9]
Regular expression not allowed to match

To verify the leap year

we use a filter row to select the rows with the month 2 and we fix the marge of the values of the day using a data validator, and then we merge all the records in a sorted merge.

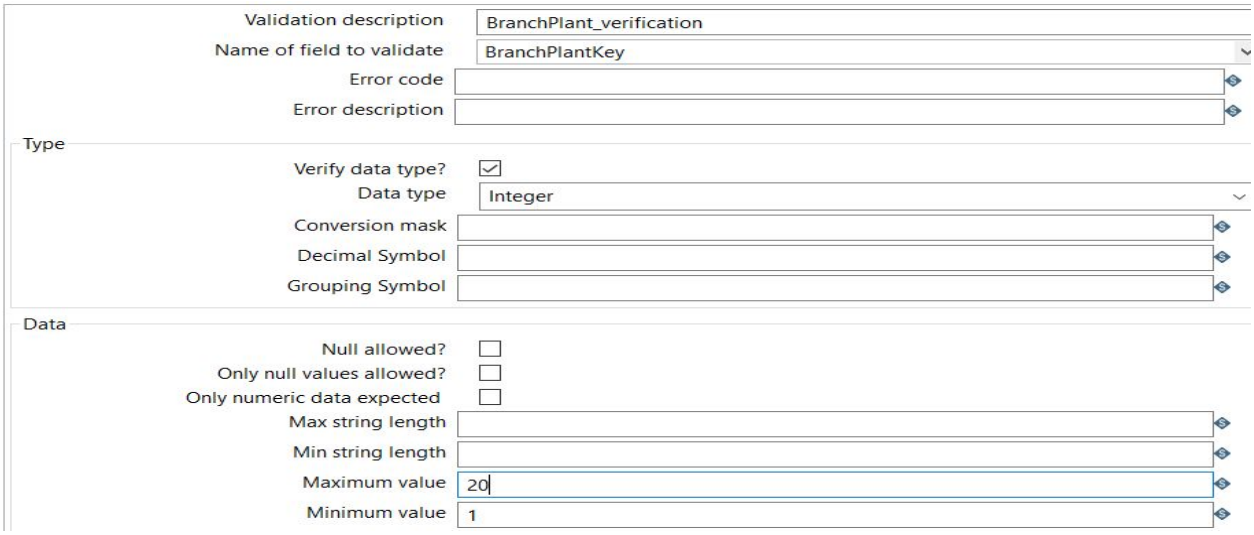






Reject invalid foreign key references: the Customer vendor key, branch plant key, transtype key, and Item master key must be valid references to rows of the respective tables in the inventory data warehouse.

BranchPlantKey	BranchPlantId	CompanyKey	CarryingCost	CostMethod	BPName
10	10	3	0.11	07	Branch Plant 10
11	11	3	0.17	07	Branch Plant 11
12	12	3	0.16	07	Branch Plant 12
13	13	4	0.05	07	Branch Plant 13
14	14	4	0.18	07	Branch Plant 14
15	15	4	0.10	07	Branch Plant 15
16	16	4	0.08	07	Branch Plant 16
17	17	5	0.21	07	Branch Plant 17
18	18	5	0.14	07	Branch Plant 18
19	19	5	0.14	07	Branch Plant 19
20	20	5	0.23	07	Branch Plant 20
NULL	NULL	NULL	NULL	NULL	NULL



ItemMasterKey	ShortItemId	SecondItemId	ThirdItemId	ItemCatCode1	ItemCatCode2	ItemDesc	UOM
10	10	Second Part 10	Thrid Part 10	3	2	Part Description 10	EA
11	11	Second Part 11	Thrid Part 11	6	7	Part Description 11	EA
12	12	Second Part 12	Thrid Part 12	6	6	Part Description 12	EA
13	13	Second Part 13	Thrid Part 13	4	3	Part Description 13	EA
14	14	Second Part 14	Thrid Part 14	3	2	Part Description 14	EA
15	15	Second Part 15	Thrid Part 15	5	8	Part Description 15	EA
16	16	Second Part 16	Thrid Part 16	6	1	Part Description 16	EA
17	17	Second Part 17	Thrid Part 17	4	4	Part Description 17	EA
18	18	Second Part 18	Thrid Part 18	7	4	Part Description 18	EA
19	19	Second Part 19	Thrid Part 19	4	5	Part Description 19	EA
20	20	Second Part 20	Thrid Part 20	3	3	Part Description 20	EA

Validation description: ItemMaster_verification

Name of field to validate: ItemMasterKey

Error code:

Error description:

Type

Verify data type? ☑

Data type: Integer

Conversion mask:

Decimal Symbol:

Grouping Symbol:

Data

Null allowed? ☐

Only null values allowed? ☐

Only numeric data expected ☐

Max string length:

Min string length:

Maximum value: 20

Minimum value: 1

TransTypeKey	TransTypeCodeId	TransDescription
1	IA	Inventory Adjustement
2	IT	Inventory Transfer
3	IS	Inventory Simple Issue
4	OV	Purchase Order Receipt
5	AR	Sales Order Shipment
NULL	NULL	NULL

Validation description	TransTypeKey_verification
Name of field to validate	CustVendorKey
Error code	
Error description	

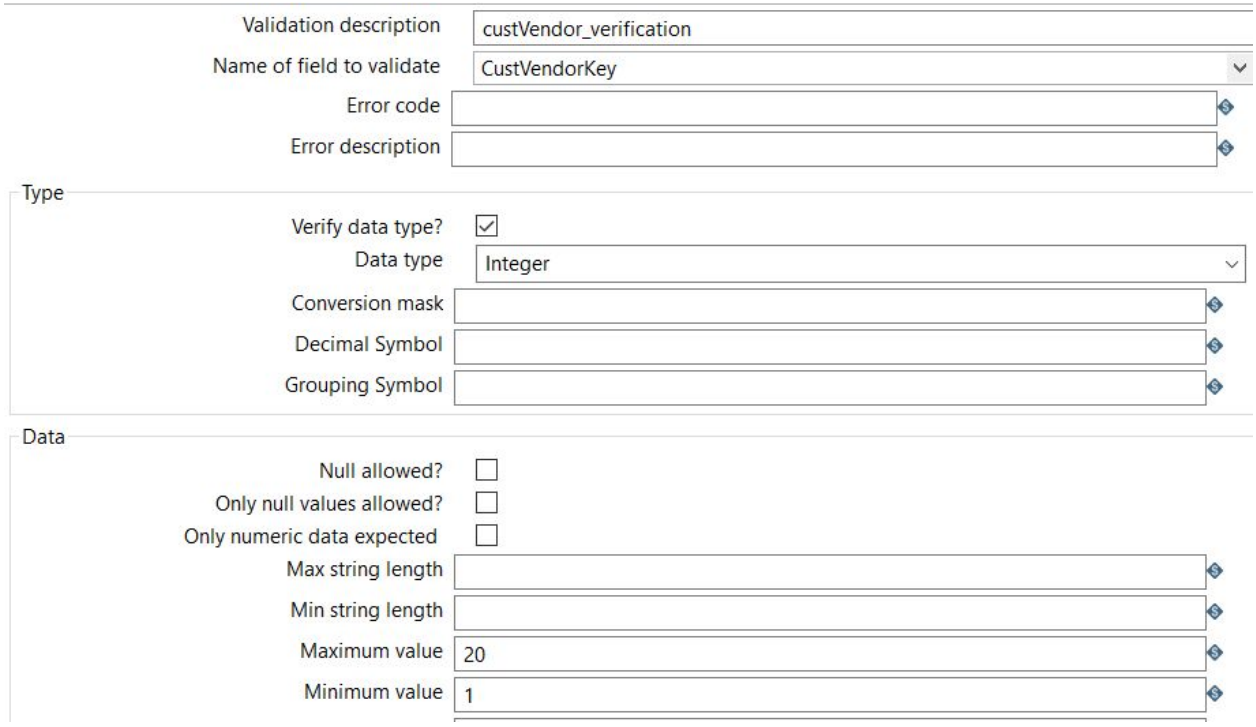
Type

Verify data type?	☑
Data type	Integer
Conversion mask	
Decimal Symbol	
Grouping Symbol	

Data

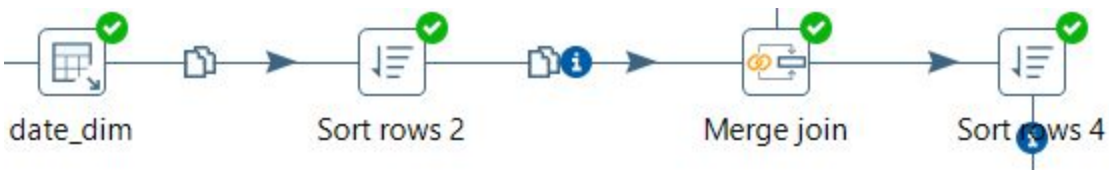
Null allowed?	☐
Only null values allowed?	☐
Only numeric data expected	☐
Max string length	
Min string length	
Maximum value	5
Minimum value	1

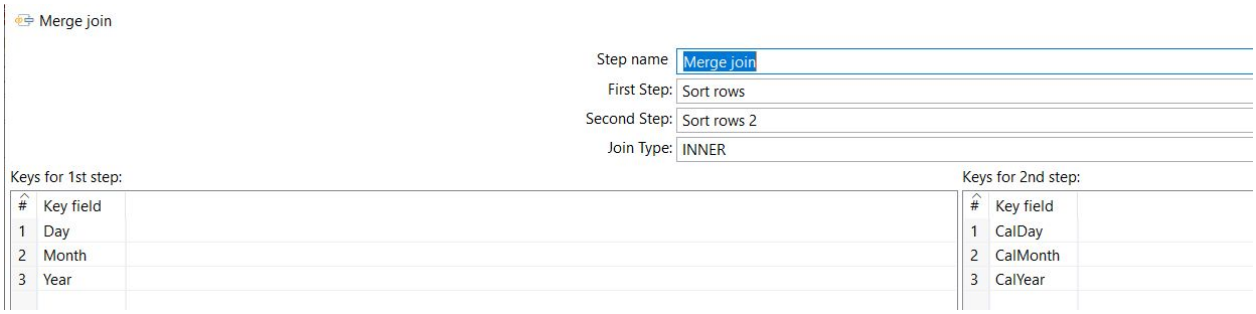
CustVendorKey	AddrBookId	Name	Address	City	State	PrimZip	Zip	Country	AddrCatCode1	AddrCatCode2
10	10	Customer 10	66683 Some Street	Cleveland	OH	44136	44136	USA	5	5
11	11	Customer 11	94803 Some Street	Boston	MA	2162	02162	USA	1	3
12	12	Customer 12	31113 Some Street	Washington	DC	20054	20054	USA	2	5
13	13	Customer 13	13112 Some Street	Miami	FL	33165	33165	USA	5	3
14	14	Customer 14	69017 Some Street	San Francisco	CA	94109	94109	USA	3	5
15	15	Customer 15	86648 Some Street	Atlanta	GA	30385	30385	USA	1	6
16	16	Customer 16	36525 Some Street	Chicago	IL	60926	60926	USA	1	5
17	17	Customer 17	72806 Some Street	New York City	NY	10231	10231	USA	4	3
18	18	Customer 18	18706 Some Street	Denver	CO	80293	80293	USA	3	1
19	19	Customer 19	14432 Some Street	Salt Lake City	UT	84194	84194	USA	5	5
20	20	Customer 20	11518 Some Street	Los Angeles	CA	90101	90101	USA	3	5
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL



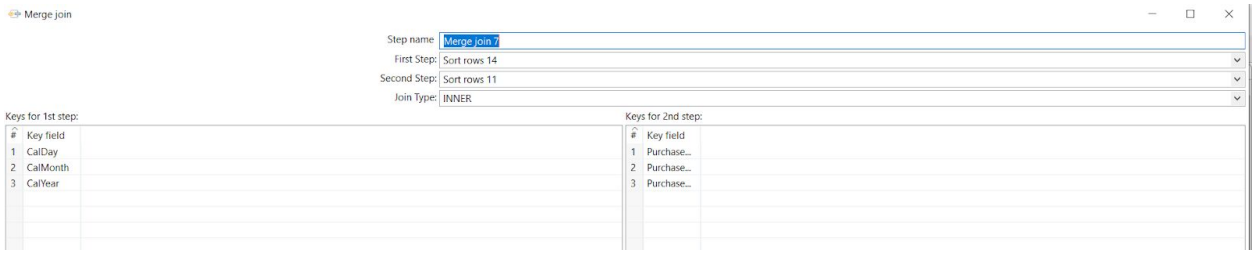
After validation, you should perform the following processing steps. These steps will enable the data to be loaded into the *Inventory_Fact* table of the inventory data warehouse.

For data source 1, the month, day, and year fields should be used to find a matching row in the *Date_Dim* table in the inventory data warehouse. After finding the matching row, the *Date_Key* value in the *Date_Dim* row should be used for the *Date_Key* value in the *Inventory_Fact* table.



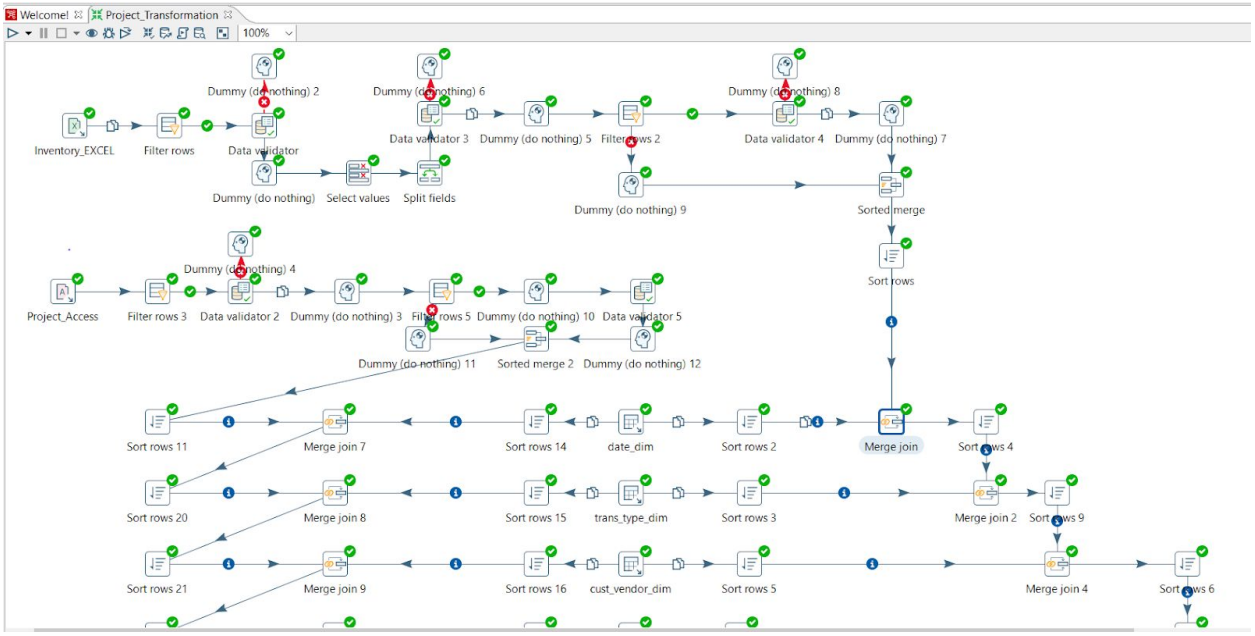


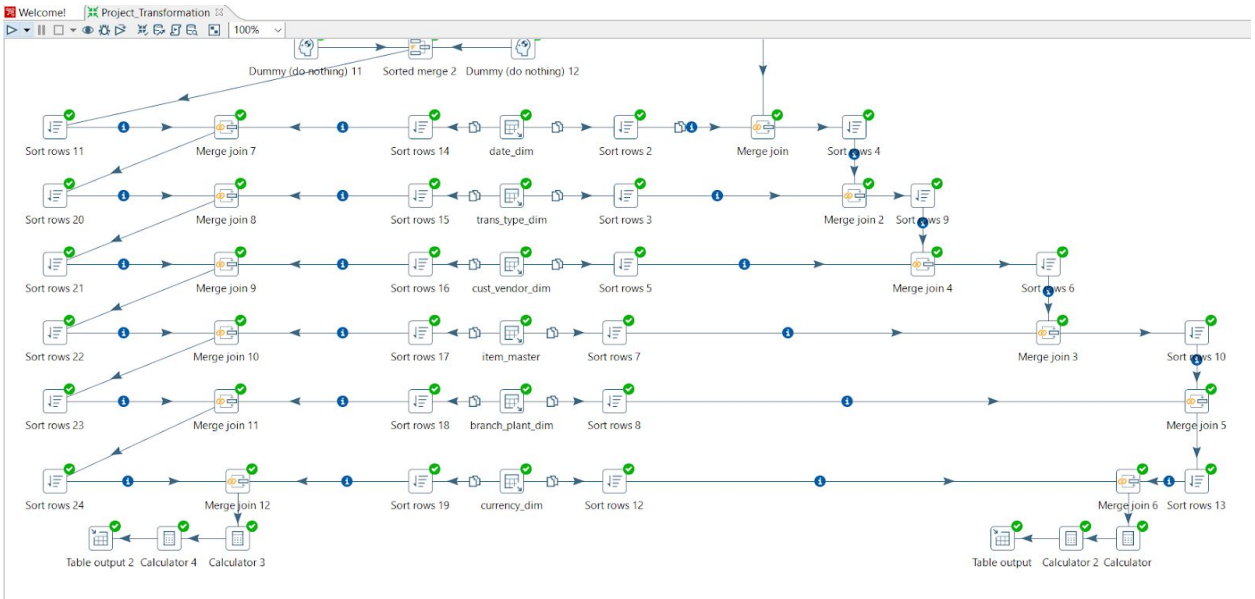
For data source 2, the Purchase Date field should be parsed into its day, month, and year components. These components should be used to find a matching row in the *Date_Dim* table. Then, the *Date_Key* value in the matching *Date_Dim* row should be used for the *Date_Key* value in the *Inventory_Fact* table. See the explanation in the following section about parsing dates in Excel data sources.



Merge Join Order

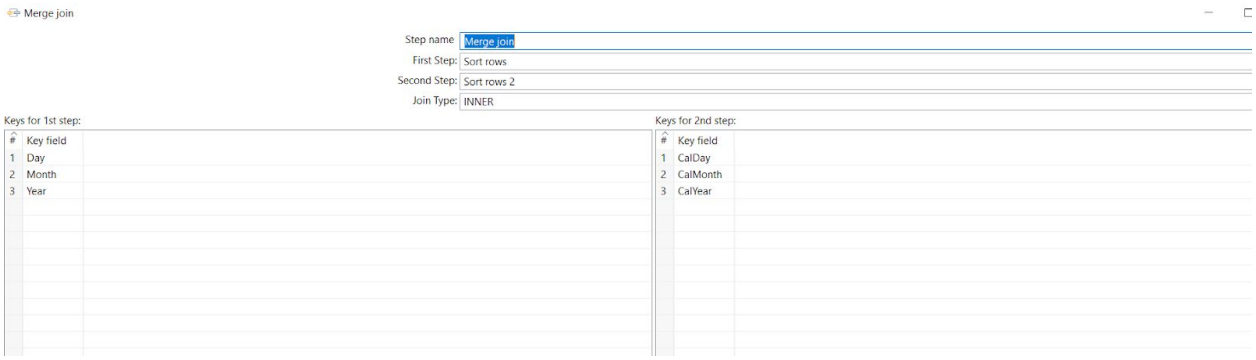
we join the tables with the two data sources in the following order .



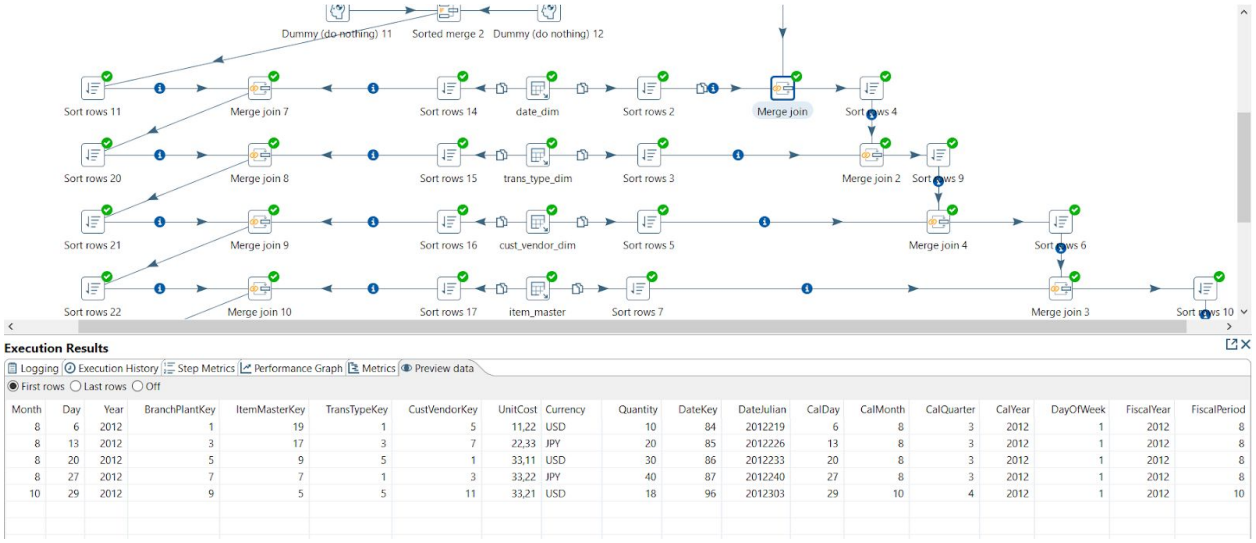


- Date_Dim

Excel source 1



After running the transformation



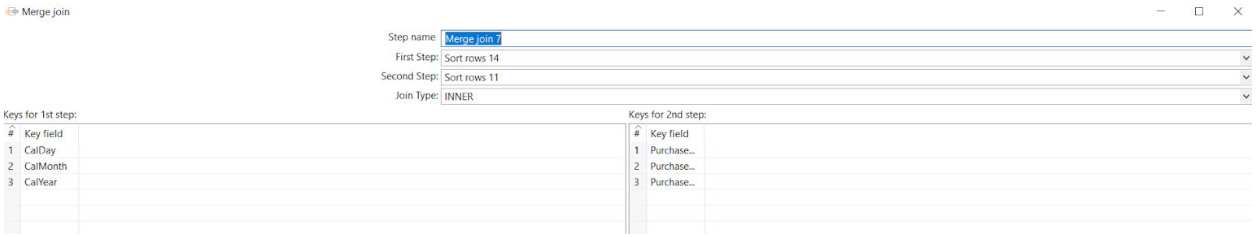
Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

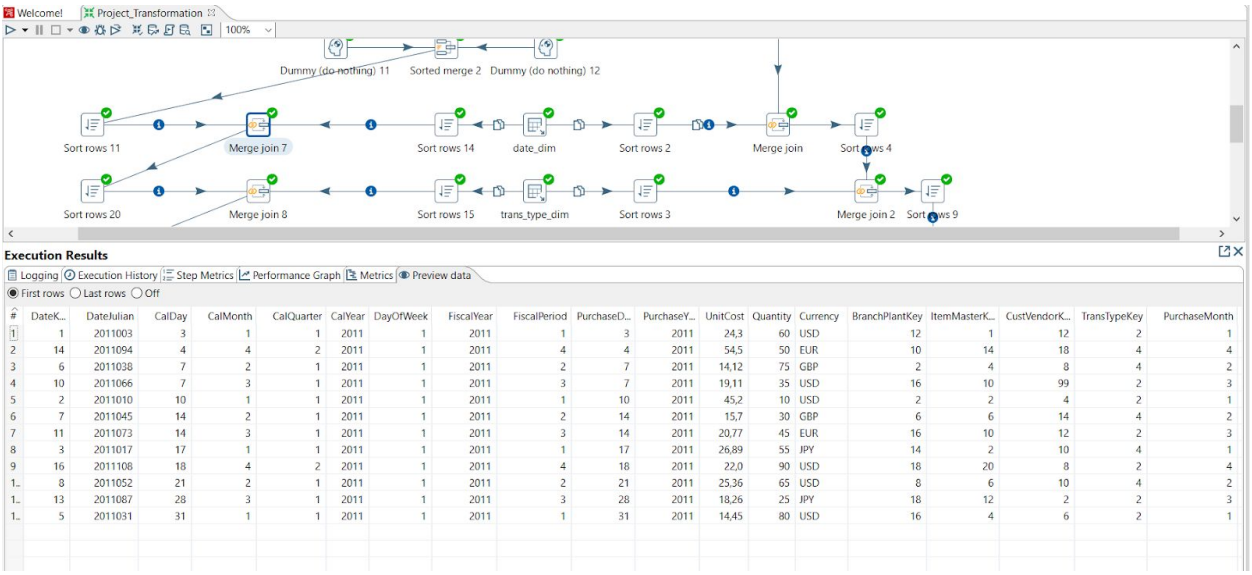
First rows | Last rows | Off

Month	Day	Year	BranchPlantKey	ItemMasterKey	TransTypeKey	CustVendorKey	UnitCost	Currency	Quantity	DateKey	DateJulian	CalDay	CalMonth	CalQuarter	CalYear	DayOfWeek	FiscalYear	FiscalPeriod
8	6	2012	1	19	1	5	11,22	USD	10	84	2012219	6	8	3	2012	1	2012	8
8	13	2012	3	17	3	7	22,33	JPY	20	85	2012226	13	8	3	2012	1	2012	8
8	20	2012	5	9	5	1	33,11	USD	30	86	2012233	20	8	3	2012	1	2012	8
8	27	2012	7	7	1	3	33,22	JPY	40	87	2012240	27	8	3	2012	1	2012	8
10	29	2012	9	5	5	11	33,21	USD	18	96	2012303	29	10	4	2012	1	2012	10

Acces source 2



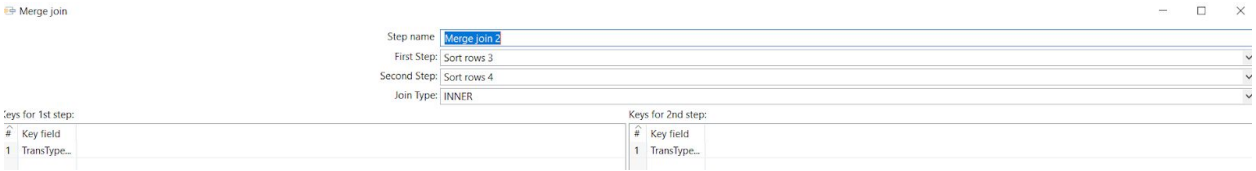
After running the transformation



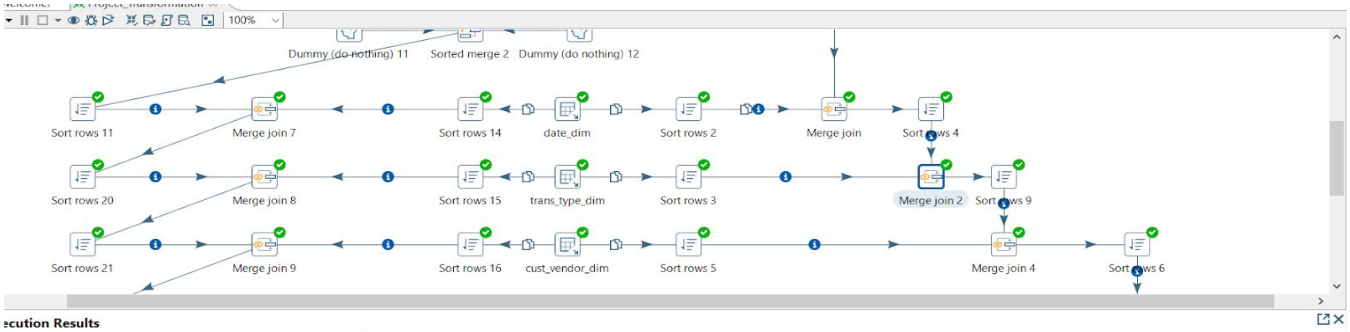
#	DateK...	DateJulian	CalDay	CalMonth	CalQuarter	CalYear	DayOfWeek	FiscalYear	FiscalPeriod	PurchaseD...	PurchaseY...	UnitCost	Quantity	Currency	BranchPlantKey	ItemMasterK...	CustVendorK...	TransTypeKey	PurchaseMonth
1	1	2011003	3	1	1	2011	1	2011	1	3	2011	24,3	60	USD	12	1	12	2	1
2	14	2011094	4	4	2	2011	1	2011	4	4	2011	54,5	50	EUR	10	14	18	4	4
3	6	2011038	7	2	1	2011	1	2011	2	7	2011	14,12	75	GBP	2	4	8	4	2
4	10	2011066	7	3	1	2011	1	2011	3	7	2011	19,11	35	USD	16	10	99	2	3
5	2	2011010	10	1	1	2011	1	2011	1	10	2011	45,2	10	USD	2	2	4	2	1
6	7	2011045	14	2	1	2011	1	2011	2	14	2011	15,7	30	GBP	6	6	14	4	2
7	11	2011073	14	3	1	2011	1	2011	3	14	2011	20,77	45	EUR	16	10	12	2	3
8	3	2011017	17	1	1	2011	1	2011	1	17	2011	26,89	55	JPY	14	2	10	4	1
9	16	2011108	18	4	2	2011	1	2011	4	18	2011	22,0	90	USD	18	20	8	2	4
1..	8	2011052	21	2	1	2011	1	2011	2	21	2011	25,36	65	USD	8	6	10	4	2
1..	13	2011087	28	3	1	2011	1	2011	3	28	2011	18,26	25	JPY	18	12	2	2	3
1..	5	2011031	31	1	1	2011	1	2011	1	31	2011	14,45	80	USD	16	4	6	2	1

- Trans_Type_Dim

Excel source 1

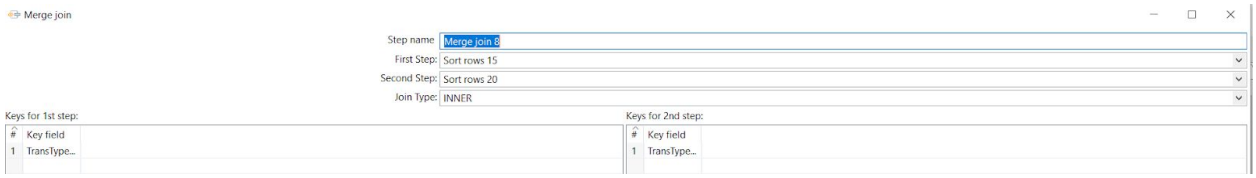


After running the transformation

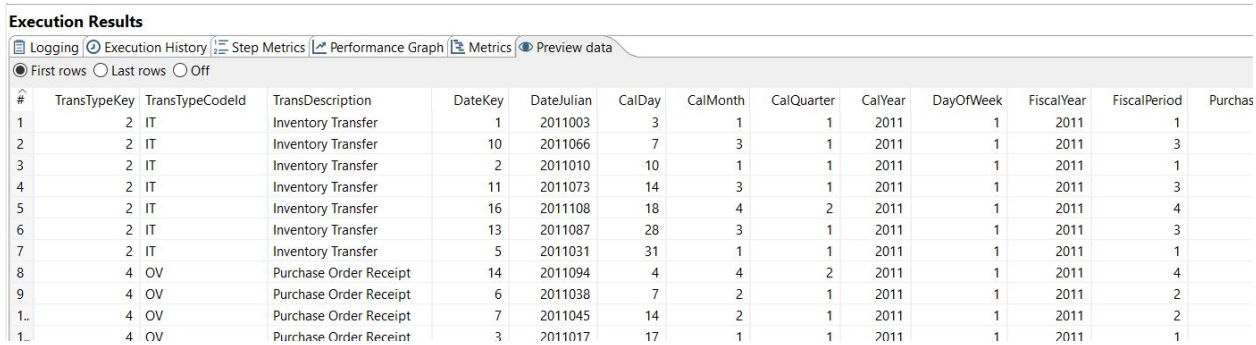


TransType...	TransTypeCo...	TransDescription	Month	Day	Year	BranchPlant...	ItemMasterKey	TransTypeKey_1	CustVendor...	UnitCost	Currency	Quantity	DateKey	DateJulian	CalDay	CalMonth	CalQuarter	CalYear	D...
1	IA	Inventory Adjus...	8	6	2012	1	19	1	5	11,22	USD	10	84	2012219	6	8	3	2012	
1	IA	Inventory Adjus...	8	27	2012	7	7	1	3	33,22	JPY	40	87	2012240	27	8	3	2012	
3	IS	Inventory Simpl...	8	13	2012	3	17	3	7	22,33	JPY	20	85	2012226	13	8	3	2012	
5	AR	Sales Order Shi...	8	20	2012	5	9	5	1	33,11	USD	30	86	2012233	20	8	3	2012	
5	AR	Sales Order Shi...	10	29	2012	9	5	5	11	33,21	USD	18	96	2012303	29	10	4	2012	

Access source 2



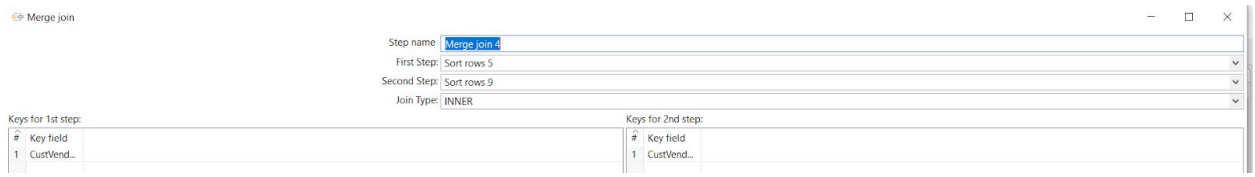
After running the transformation



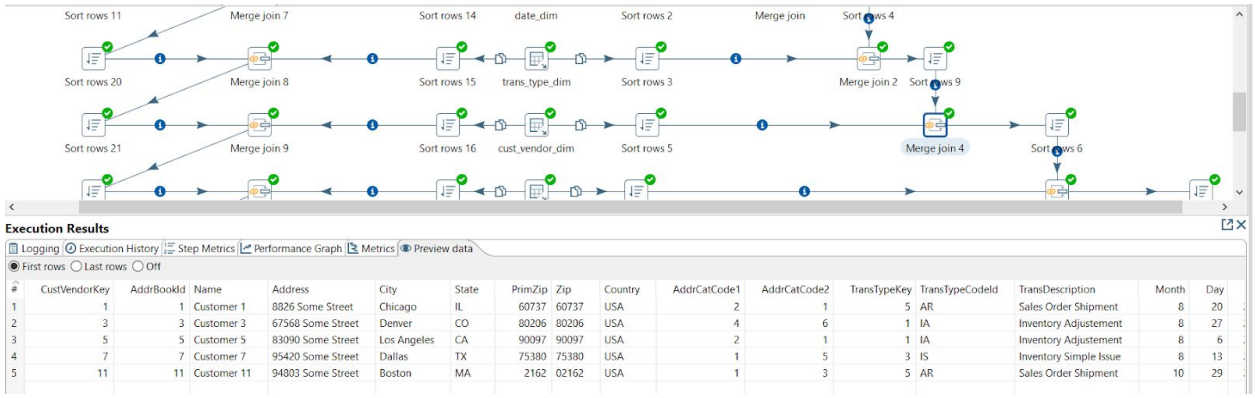
#	TransTypeKey	TransTypeCodeId	TransDescription	DateKey	DateJulian	CalDay	CalMonth	CalQuarter	CalYear	DayOfWeek	FiscalYear	FiscalPeriod	Purchas
1	2	IT	Inventory Transfer	1	2011003	3	1	1	2011	1	2011	1	
2	2	IT	Inventory Transfer	10	2011066	7	3	1	2011	1	2011	3	
3	2	IT	Inventory Transfer	2	2011010	10	1	1	2011	1	2011	1	
4	2	IT	Inventory Transfer	11	2011073	14	3	1	2011	1	2011	3	
5	2	IT	Inventory Transfer	16	2011108	18	4	2	2011	1	2011	4	
6	2	IT	Inventory Transfer	13	2011087	28	3	1	2011	1	2011	3	
7	2	IT	Inventory Transfer	5	2011031	31	1	1	2011	1	2011	1	
8	4	OV	Purchase Order Receipt	14	2011094	4	4	2	2011	1	2011	4	
9	4	OV	Purchase Order Receipt	6	2011038	7	2	1	2011	1	2011	2	
1..	4	OV	Purchase Order Receipt	7	2011045	14	2	1	2011	1	2011	2	
1..	4	OV	Purchase Order Receipt	3	2011017	17	1	1	2011	1	2011	1	

·Cust_Vendor_Dim

Excel source 1

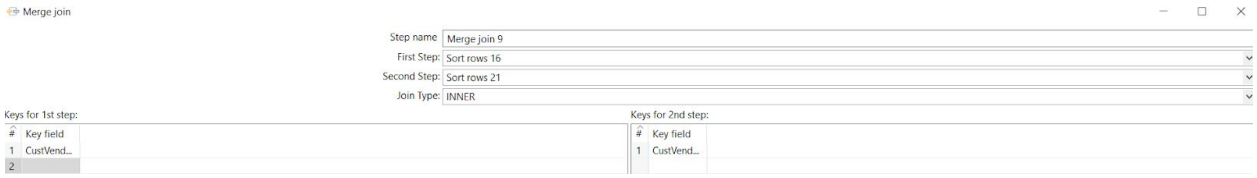


After running the transformation

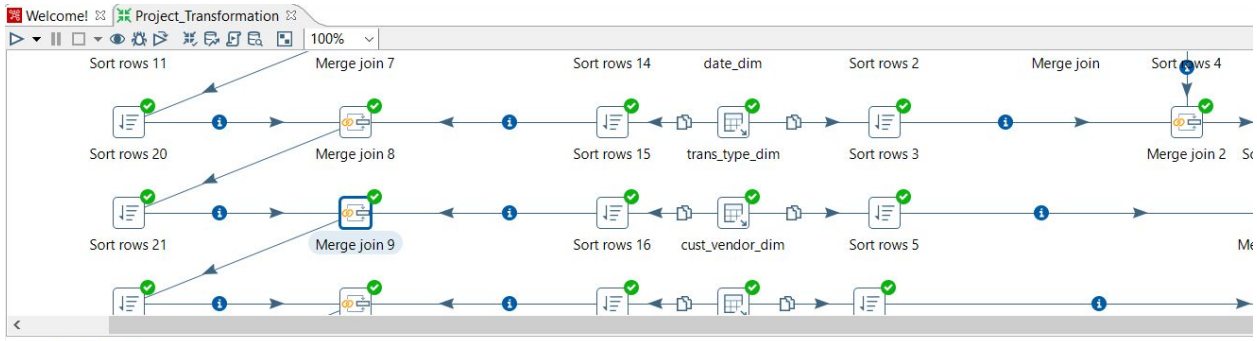


#	CustVendorKey	AddrBookId	Name	Address	City	State	PrimZip	Zip	Country	AddrCatCode1	AddrCatCode2	TransTypeKey	TransTypeCodeId	TransDescription	Month	Day
1	1	1	Customer 1	8826 Some Street	Chicago	IL	60737	60737	USA	2	1	5	AR	Sales Order Shipment	8	20
2	3	3	Customer 3	67568 Some Street	Denver	CO	80206	80206	USA	4	6	1	IA	Inventory Adjustement	8	27
3	5	5	Customer 5	83090 Some Street	Los Angeles	CA	90097	90097	USA	2	1	1	IA	Inventory Adjustement	8	6
4	7	7	Customer 7	95420 Some Street	Dallas	TX	75380	75380	USA	1	5	3	IS	Inventory Simple Issue	8	13
5	11	11	Customer 11	94803 Some Street	Boston	MA	2162	02162	USA	1	3	5	AR	Sales Order Shipment	10	29

Access source 2



After running the transformation



Execution Results

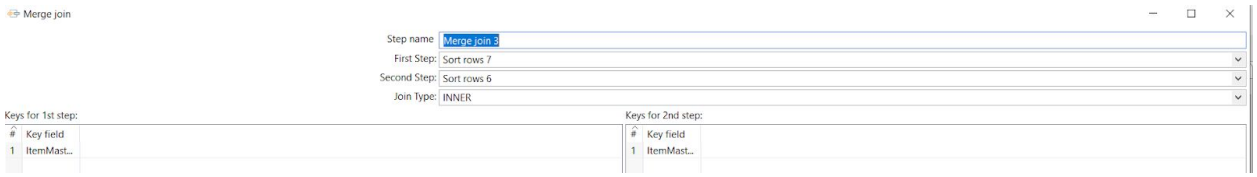
Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

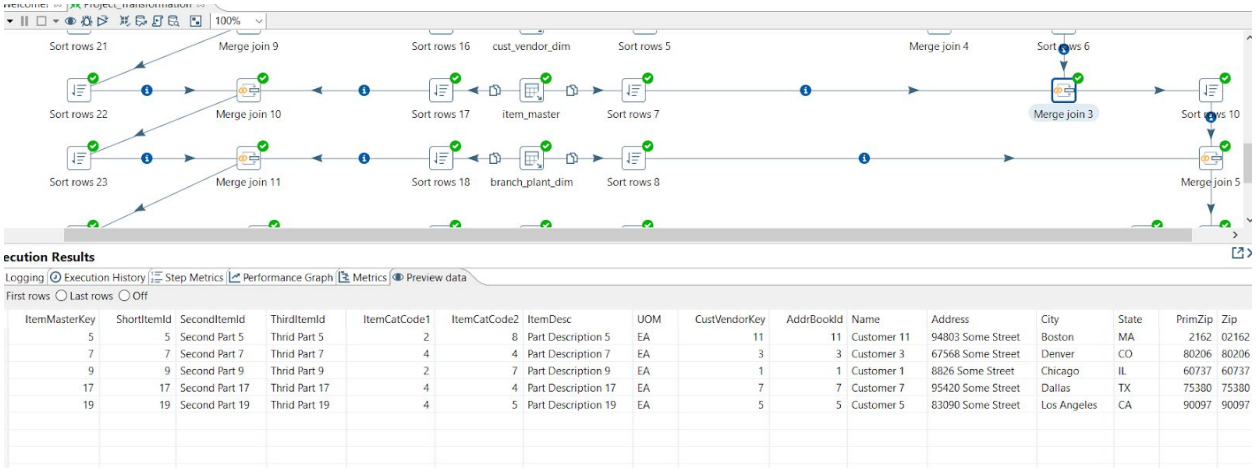
#	CustVendorKey	AddrBookId	Name	Address	City	State	PrimZip	Zip	Country	AddrCatCode1	AddrCatCode2	TransTypeKey
1	2	2	Customer 2	56647 Some Street	New York City	NY	10251	10251	USA	4	5	2
2	4	4	Customer 4	85686 Some Street	Salt Lake City	UT	84128	84128	USA	1	6	2
3	6	6	Customer 6	42055 Some Street	Houston	TX	77112	77112	USA	2	1	2
4	8	8	Customer 8	59511 Some Street	Seattle	WA	98292	98292	USA	5	6	2
5	8	8	Customer 8	59511 Some Street	Seattle	WA	98292	98292	USA	5	6	4
6	10	10	Customer 10	66683 Some Street	Cleveland	OH	44136	44136	USA	5	5	4
7	10	10	Customer 10	66683 Some Street	Cleveland	OH	44136	44136	USA	5	5	4
8	12	12	Customer 12	31113 Some Street	Washington	DC	20054	20054	USA	2	5	2
9	12	12	Customer 12	31113 Some Street	Washington	DC	20054	20054	USA	2	5	2
1..	14	14	Customer 14	69017 Some Street	San Francisco	CA	94109	94109	USA	3	5	4
1..	18	18	Customer 18	18706 Some Street	Denver	CO	80293	80293	USA	3	1	4

· Item_Master_Dim

Excel source 1

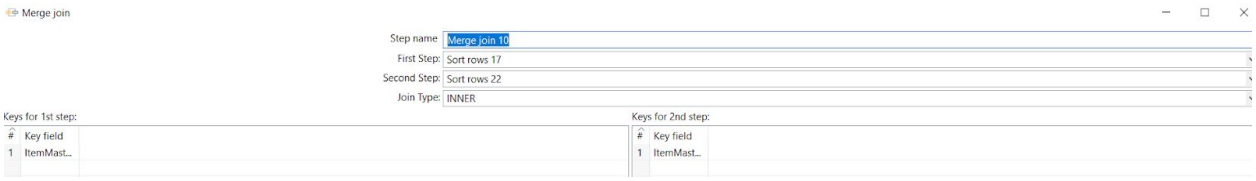


After running the transformation

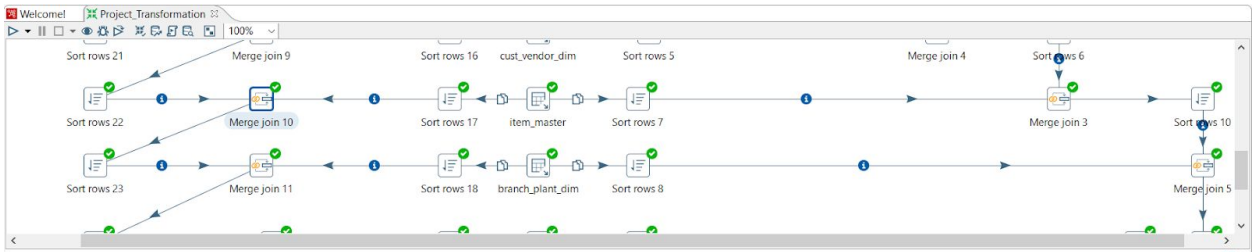


ItemMasterKey	ShortItemId	SecondItemId	ThirdItemId	ItemCatCode1	ItemCatCode2	ItemDesc	UOM	CustVendorKey	AddrBookId	Name	Address	City	State	PrimZip	Zip
5	5	Second Part 5	Thrid Part 5	2	8	Part Description 5	EA	11	11	Customer 11	94803 Some Street	Boston	MA	2162	02162
7	7	Second Part 7	Thrid Part 7	4	4	Part Description 7	EA	3	3	Customer 3	67568 Some Street	Denver	CO	80206	80206
9	9	Second Part 9	Thrid Part 9	2	7	Part Description 9	EA	1	1	Customer 1	8826 Some Street	Chicago	IL	60737	60737
17	17	Second Part 17	Thrid Part 17	4	4	Part Description 17	EA	7	7	Customer 7	95420 Some Street	Dallas	TX	75380	75380
19	19	Second Part 19	Thrid Part 19	4	5	Part Description 19	EA	5	5	Customer 5	83090 Some Street	Los Angeles	CA	90097	90097

Access source 2



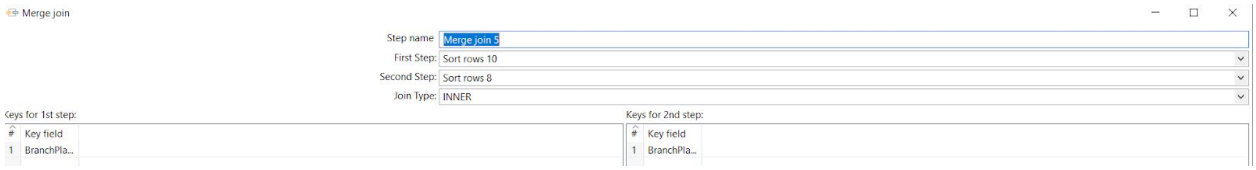
After running the transformation



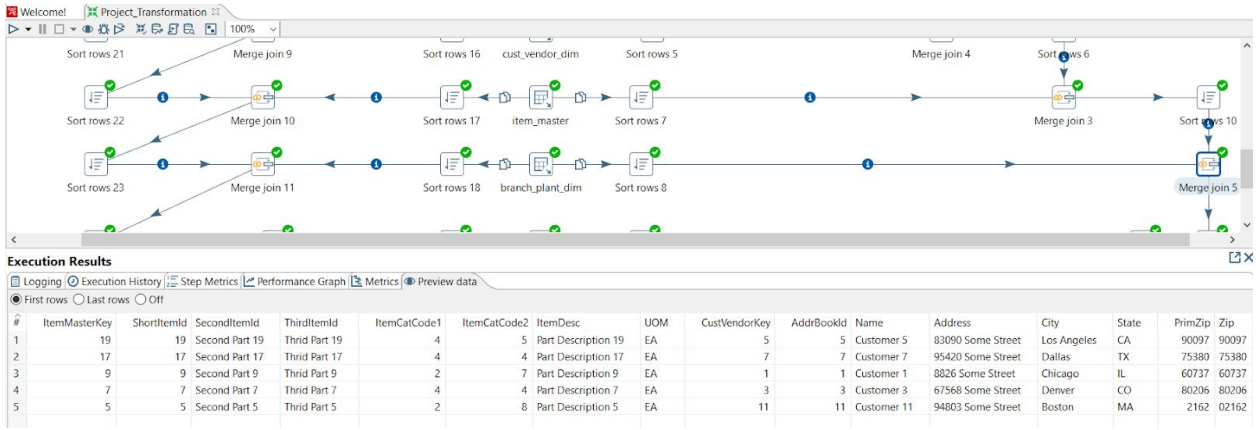
#	ItemMasterKey	ShortItemId	SecondItemId	ThirdItemId	ItemCatCode1	ItemCatCode2	ItemDesc	UOM	CustVendorKey	AddrBookId	Name	Address	City	State	PrimZip	Zip
1	1	1	Second Part 1	Thrid Part 1	7	5	Part Description 1	EA	12	12	Customer 12	31113 Some Street	Washington	DC	20054	2005
2	2	2	Second Part 2	Thrid Part 2	2	7	Part Description 2	EA	4	4	Customer 4	85686 Some Street	Salt Lake City	UT	84128	8412
3	2	2	Second Part 2	Thrid Part 2	2	7	Part Description 2	EA	10	10	Customer 10	66683 Some Street	Cleveland	OH	44136	4413
4	4	4	Second Part 4	Thrid Part 4	1	4	Part Description 4	EA	6	6	Customer 6	42055 Some Street	Houston	TX	77112	7711
5	4	4	Second Part 4	Thrid Part 4	1	4	Part Description 4	EA	8	8	Customer 8	59511 Some Street	Seattle	WA	98292	9829
6	6	6	Second Part 6	Thrid Part 6	2	2	Part Description 6	EA	10	10	Customer 10	66683 Some Street	Cleveland	OH	44136	4413
7	6	6	Second Part 6	Thrid Part 6	2	2	Part Description 6	EA	14	14	Customer 14	69017 Some Street	San Francisco	CA	94109	9410
8	10	10	Second Part 10	Thrid Part 10	3	2	Part Description 10	EA	12	12	Customer 12	31113 Some Street	Washington	DC	20054	2005
9	12	12	Second Part 12	Thrid Part 12	6	6	Part Description 12	EA	2	2	Customer 2	56647 Some Street	New York City	NY	10251	1025
1..	14	14	Second Part 14	Thrid Part 14	3	2	Part Description 14	EA	18	18	Customer 18	18706 Some Street	Denver	CO	80293	8029
1..	20	20	Second Part 20	Thrid Part 20	3	3	Part Description 20	EA	8	8	Customer 8	59511 Some Street	Seattle	WA	98292	9829

· Branch_Plant_Dim

Excel source 1

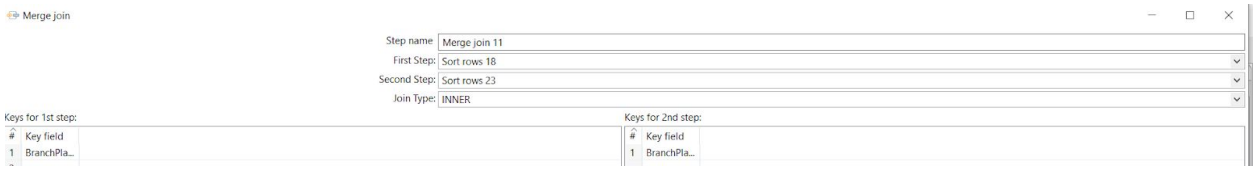


After running the transformation

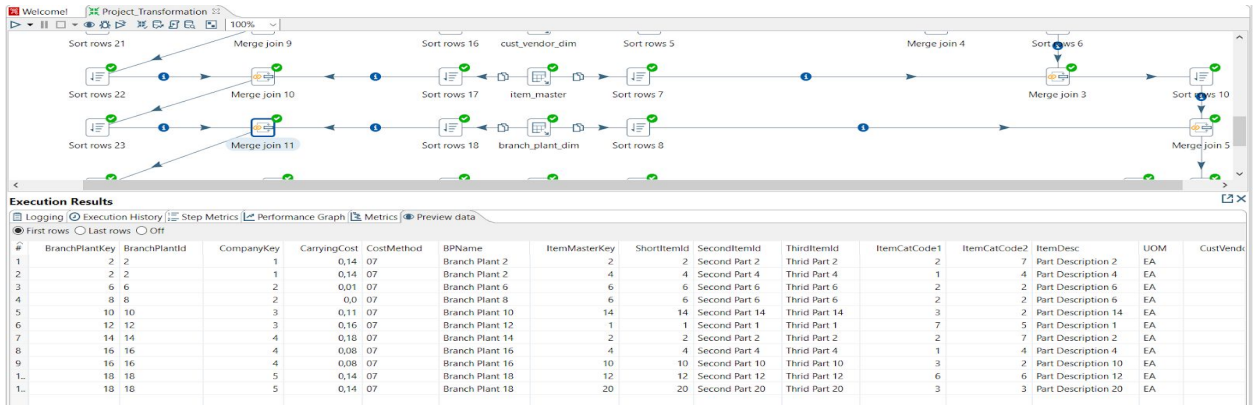


#	ItemMasterKey	ShortItemId	SecondItemId	ThirdItemId	ItemCatCode1	ItemCatCode2	ItemDesc	UOM	CustVendorKey	AddrBookId	Name	Address	City	State	PrimZip	Zip
1	19	19	Second Part 19	Thrid Part 19	4	5	Part Description 19	EA	5	5	Customer 5	83090 Some Street	Los Angeles	CA	90097	90097
2	17	17	Second Part 17	Thrid Part 17	4	4	Part Description 17	EA	7	7	Customer 7	95420 Some Street	Dallas	TX	75380	75380
3	9	9	Second Part 9	Thrid Part 9	2	7	Part Description 9	EA	1	1	Customer 1	8826 Some Street	Chicago	IL	60737	60737
4	7	7	Second Part 7	Thrid Part 7	4	4	Part Description 7	EA	3	3	Customer 3	67568 Some Street	Denver	CO	80206	80206
5	5	5	Second Part 5	Thrid Part 5	2	8	Part Description 5	EA	11	11	Customer 11	94803 Some Street	Boston	MA	2162	02162

Access source 2



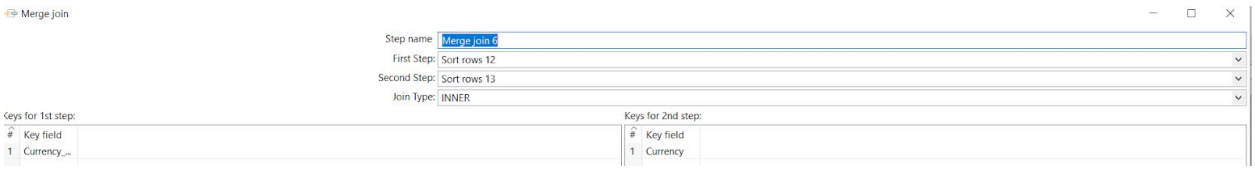
After running the transformation



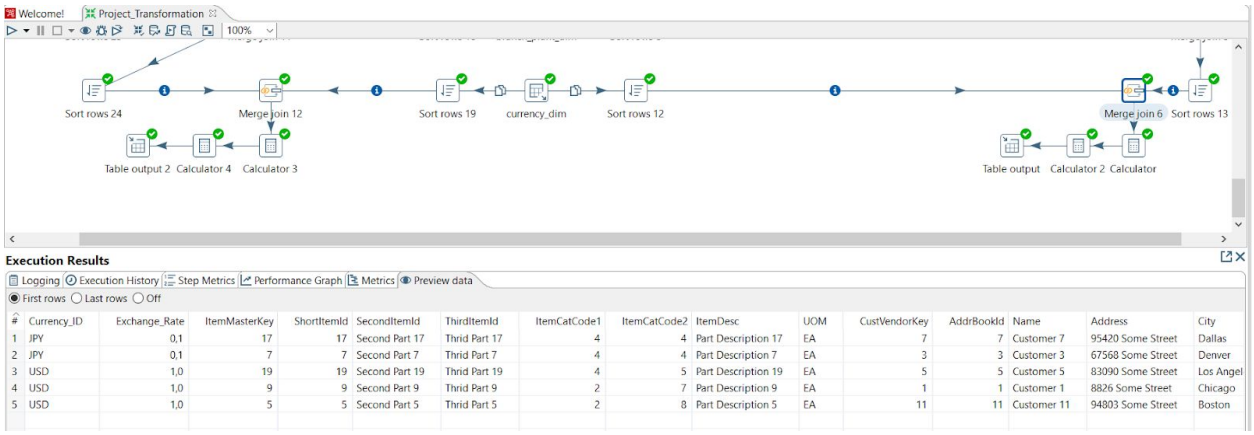
#	BranchPlantKey	BranchPlantId	CompanyKey	CarryingCost	CostMethod	BPName	ItemMasterKey	ShortItemId	SecondItemId	ThirdItemId	ItemCatCode1	ItemCatCode2	ItemDesc	UOM	CustVendo
1	2	2	1	0,14	07	Branch Plant 2	2	2	Second Part 2	Thrid Part 2	2	7	Part Description 2	EA	
2	2	2	1	0,14	07	Branch Plant 2	4	4	Second Part 4	Thrid Part 4	1	4	Part Description 4	EA	
3	6	6	2	0,01	07	Branch Plant 6	6	6	Second Part 6	Thrid Part 6	2	2	Part Description 6	EA	
4	8	8	2	0,0	07	Branch Plant 8	6	6	Second Part 6	Thrid Part 6	2	2	Part Description 6	EA	
5	10	10	3	0,11	07	Branch Plant 10	14	14	Second Part 14	Thrid Part 14	3	2	Part Description 14	EA	
6	12	12	3	0,16	07	Branch Plant 12	1	1	Second Part 1	Thrid Part 1	7	5	Part Description 1	EA	
7	14	14	4	0,18	07	Branch Plant 14	2	2	Second Part 2	Thrid Part 2	2	7	Part Description 2	EA	
8	16	16	4	0,08	07	Branch Plant 16	4	4	Second Part 4	Thrid Part 4	1	4	Part Description 4	EA	
9	16	16	4	0,08	07	Branch Plant 16	10	10	Second Part 10	Thrid Part 10	3	2	Part Description 10	EA	
1..	18	18	5	0,14	07	Branch Plant 18	12	12	Second Part 12	Thrid Part 12	6	6	Part Description 12	EA	
1..	18	18	5	0,14	07	Branch Plant 18	20	20	Second Part 20	Thrid Part 20	3	3	Part Description 20	EA	

· Currency_Dim

Excel source 1

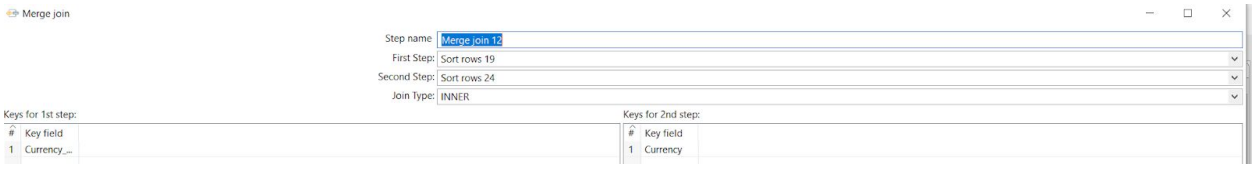


After running the transformation

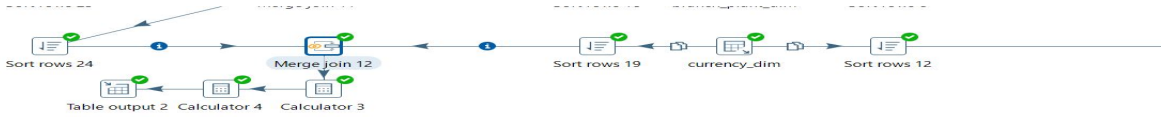


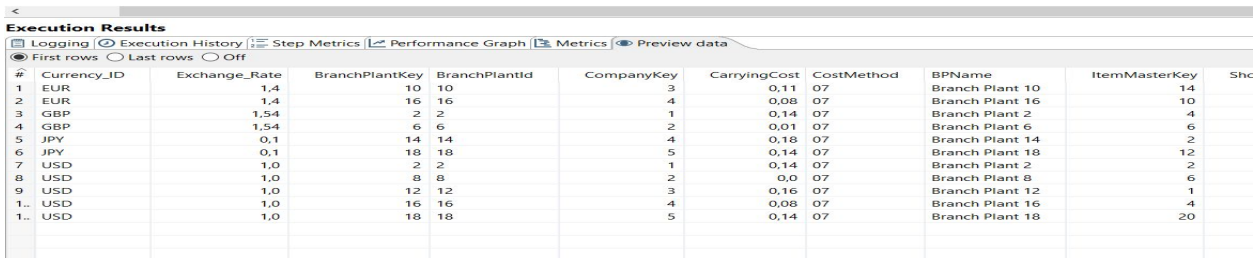
#	Currency_ID	Exchange_Rate	ItemMasterKey	ShortItemId	SecondItemId	ThirdItemId	ItemCatCode1	ItemCatCode2	ItemDesc	UOM	CustVendorKey	AddrBookId	Name	Address	City
1	JPY	0,1	17	17	Second Part 17	Thrid Part 17	4	4	Part Description 17	EA	7	7	Customer 7	95420 Some Street	Dallas
2	JPY	0,1	7	7	Second Part 7	Thrid Part 7	4	4	Part Description 7	EA	3	3	Customer 3	67568 Some Street	Denver
3	USD	1,0	19	19	Second Part 19	Thrid Part 19	4	5	Part Description 19	EA	5	5	Customer 5	83090 Some Street	Los Angel
4	USD	1,0	9	9	Second Part 9	Thrid Part 9	2	7	Part Description 9	EA	1	1	Customer 1	8826 Some Street	Chicago
5	USD	1,0	5	5	Second Part 5	Thrid Part 5	2	8	Part Description 5	EA	11	11	Customer 11	94803 Some Street	Boston

Access source 2



After running the transformation

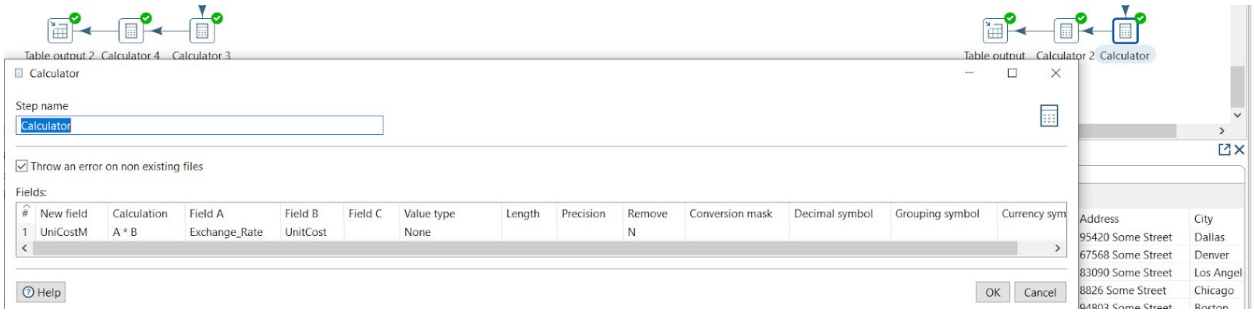




#	Currency_ID	Exchange_Rate	BranchPlantKey	BranchPlantId	CompanyKey	CarryingCost	CostMethod	BPName	ItemMasterKey	Sh
1	EUR	1,4	10	10	3	0,11	07	Branch Plant 10	14	
2	EUR	1,4	16	16	4	0,08	07	Branch Plant 16	10	
3	GBP	1,54	2	2	1	0,14	07	Branch Plant 2	4	
4	GBP	1,54	6	6	2	0,01	07	Branch Plant 6	6	
5	JPY	0,1	14	14	4	0,18	07	Branch Plant 14	2	
6	JPY	0,1	18	18	5	0,14	07	Branch Plant 18	12	
7	USD	1,0	2	2	1	0,14	07	Branch Plant 2	2	
8	USD	1,0	8	8	2	0,0	07	Branch Plant 8	6	
9	USD	1,0	12	12	3	0,16	07	Branch Plant 12	1	
1..	USD	1,0	16	16	4	0,08	07	Branch Plant 16	4	
1..	USD	1,0	18	18	5	0,14	07	Branch Plant 18	20	

The *UnitCost* column in the *Inventory_Fact* table is computed as the currency conversion factor times the Unit Cost field in the data source.

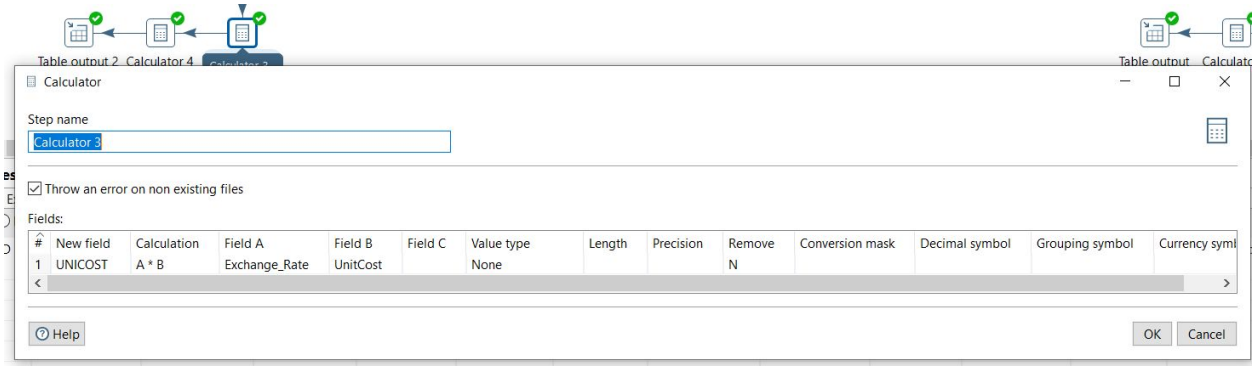
Excel source 1



After running the transformation

Currency_ID	Exchange_Rate	UnitCost	UniCostM
JPY	0,1	22,33	2,233
JPY	0,1	33,22	3,322
USD	1,0	11,22	11,22
USD	1,0	33,11	33,11
USD	1,0	33,21	33,21

Access source 2

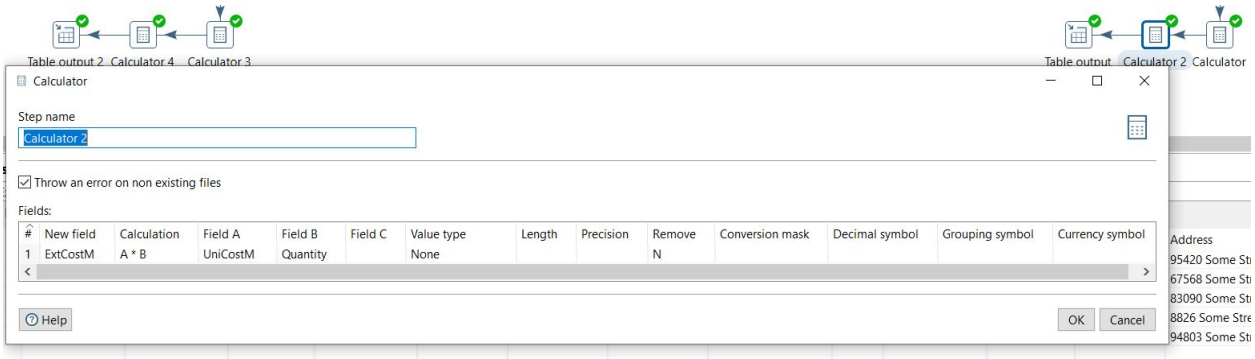


After running the transformation

Currency_ID	Exchange_Rate	UnitCost	UNICOST
EUR	1,4	54,5	76,3
EUR	1,4	20,77	29,078
GBP	1,54	14,12	21,7448
GBP	1,54	15,7	24,178
JPY	0,1	26,89	2,689
JPY	0,1	18,26	1,826
USD	1,0	45,2	45,2
USD	1,0	25,36	25,36
USD	1,0	24,3	24,3
USD	1,0	14,45	14,45
USD	1,0	22,0	22,0

The *ExtCost* column in the *Inventory_Fact* table is computed as the Unit Cost (after currency conversion) times the Quantity. We need to use a Calculator step.

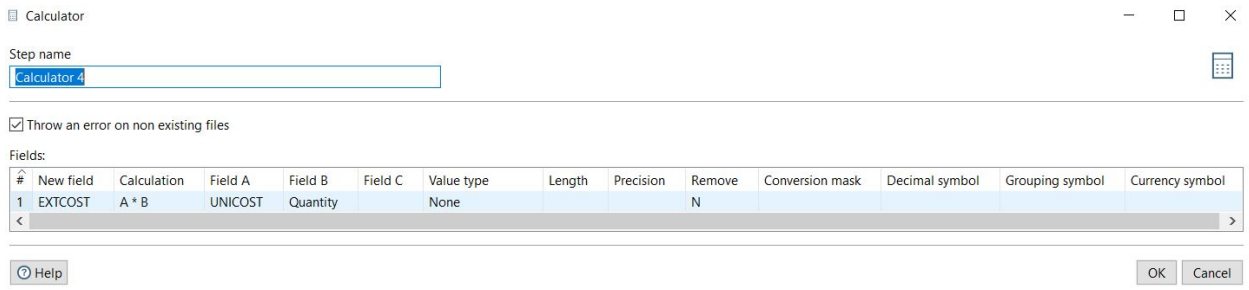
Excel source 1



After running the transformation

Currency	Quantity	UniCostM	ExtCostM
JPY	20	2,233	44,66
JPY	40	3,322	132,88
USD	10	11,22	112,2
USD	30	33,11	993,3
USD	18	33,21	597,78

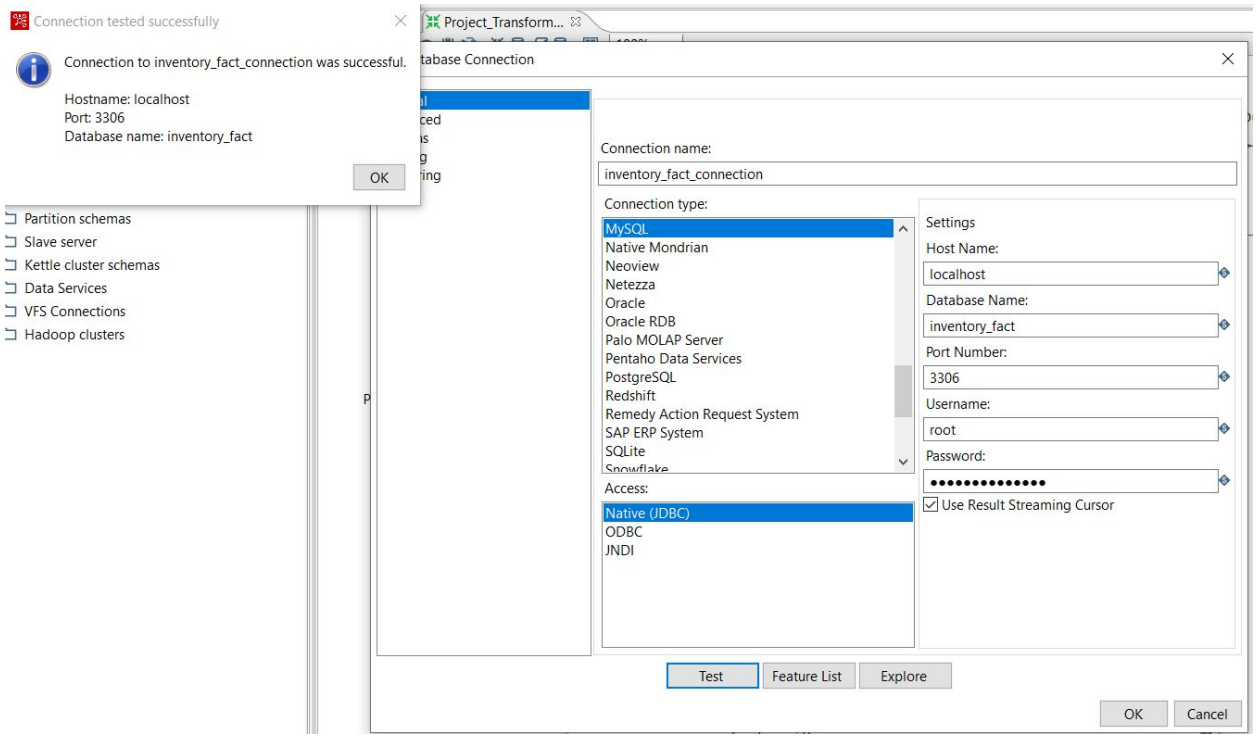
Access source 2



After running the transformation

Quantity	Currency	BranchPlantKey_1	ItemMasterKey_1	CustVendorKey_1	TransTypeKey_1	PurchaseMonth	UNICOST	EXTCOST
50	EUR	10	14	18	4	4	76,3	3815,0
45	EUR	16	10	12	2	3	29,078	1308,51
75	GBP	2	4	8	4	2	21,7448	1630,86
30	GBP	6	6	14	4	2	24,178	725,34
55	JPY	14	2	10	4	1	2,689	147,895
25	JPY	18	12	2	2	3	1,826	45,65
10	USD	2	2	4	2	1	45,2	452,0
65	USD	8	6	10	4	2	25,36	1648,4
60	USD	12	1	12	2	1	24,3	1458,0
80	USD	16	4	6	2	1	14,45	1156,0
90	USD	18	20	8	2	4	22,0	1980,0

Creation of database connection

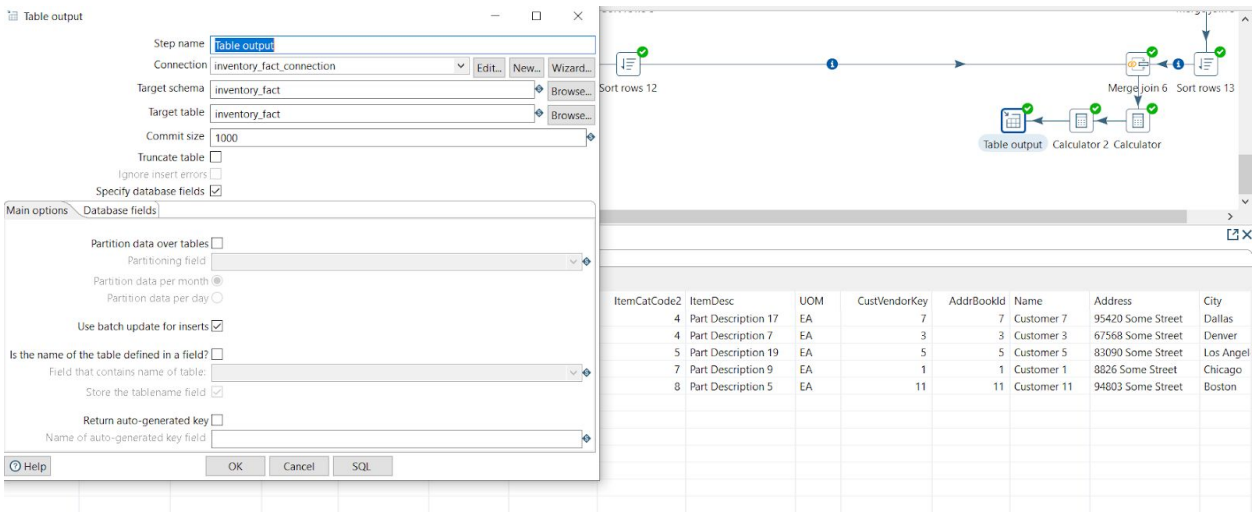


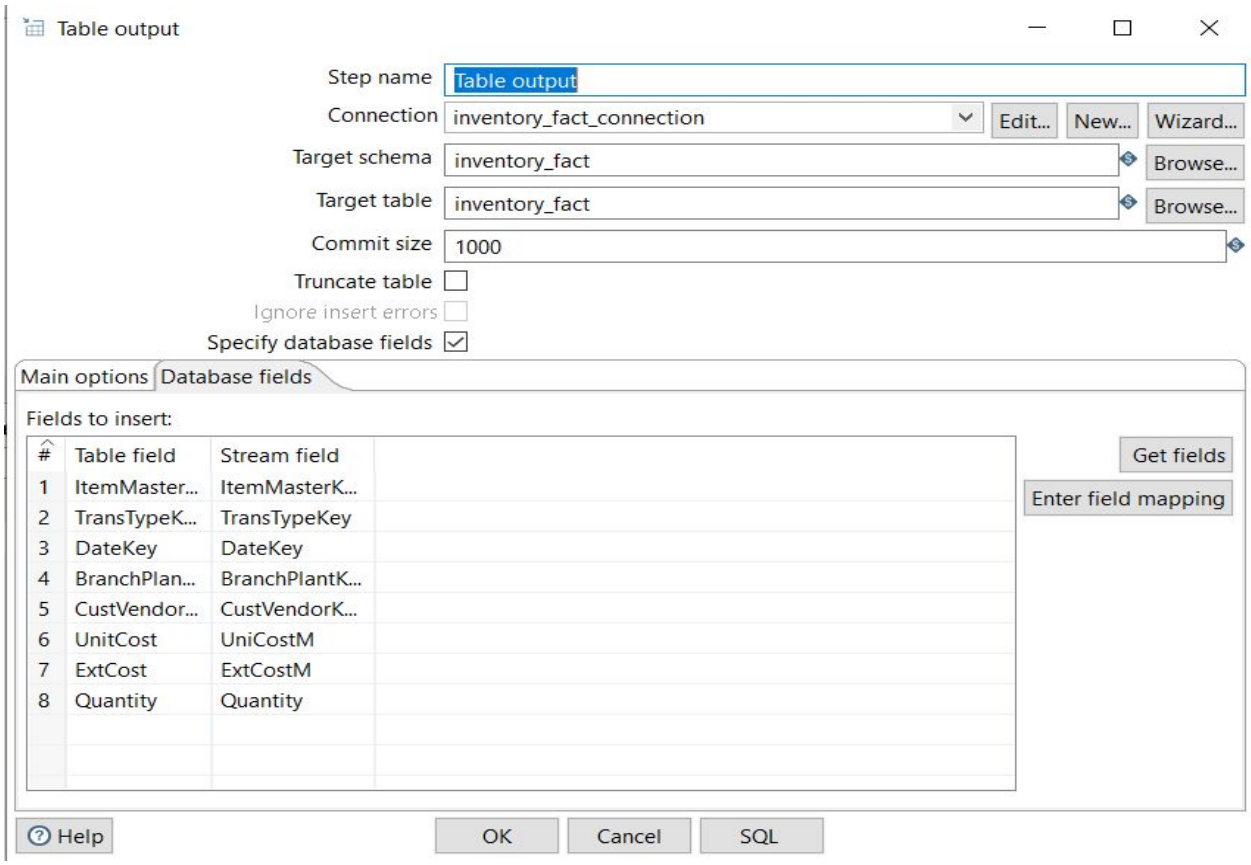
the limit line of the original records in the data warehouse created previously.

InventoryKey	DateKey	ItemMasterKey	TransTypeKey	BranchPlantKey	CustVendorKey	UnitCost	Quantity	ExtCost
18286	88	19	4	6	18	915.0000	9528.0000	8718120.00
18287	20	2	4	16	5	1310.0000	10884.0000	14258040.00
18288	37	4	4	2	2	1342.0000	8250.0000	11071500.00
18289	73	14	4	7	1	772.0000	9626.0000	7431272.00
18290	70	6	4	18	18	628.0000	10734.0000	6740952.00
18291	64	20	4	17	5	599.0000	14737.0000	8827463.00
18292	9	11	4	3	13	1201.0000	6874.0000	8255674.00
18293	4	1	4	20	12	604.0000	14153.0000	8548412.00
18294	66	15	4	20	11	941.0000	9224.0000	8679784.00
18295	82	5	4	4	16	1108.0000	11692.0000	12954736.00
18296	52	15	4	11	10	620.0000	13557.0000	8405340.00
18297	88	10	4	14	15	1036.0000	5396.0000	5590256.00
18298	50	11	4	13	15	961.0000	12025.0000	11556025.00
18299	66	6	4	8	10	919.0000	11717.0000	10767923.00
18300	2	5	4	11	6	1458.0000	13855.0000	20200590.00
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

To insert the data extracted ,verified, transformed in the database, we use the table output step

Excel source 1



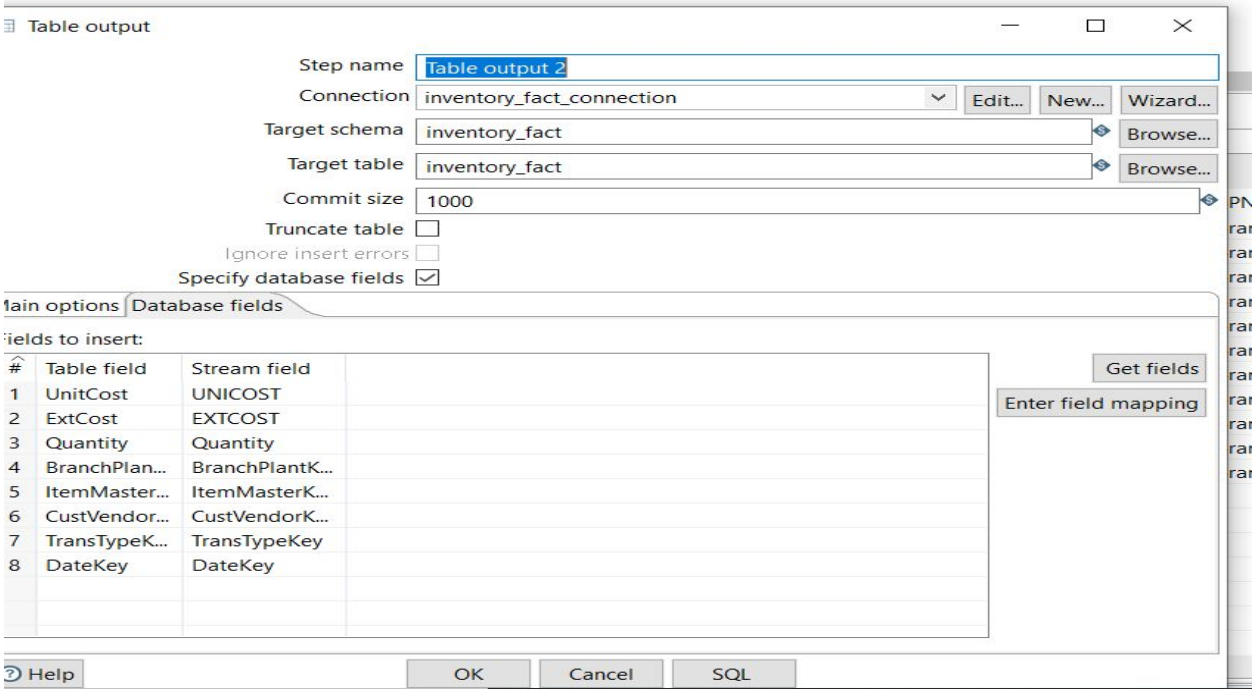


After running the transformation

5 rows added

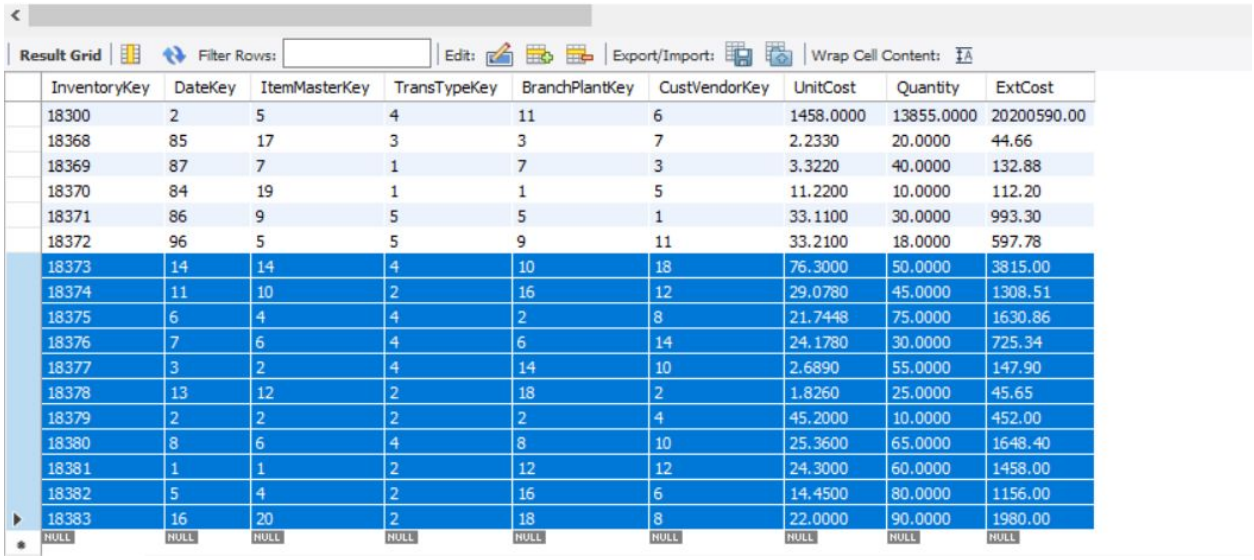
InventoryKey	DateKey	ItemMasterKey	TransTypeKey	BranchPlantKey	CustVendorKey	UnitCost	Quantity	ExtCost
18289	73	14	4	7	1	772.0000	9626.0000	7431272.00
18290	70	6	4	18	18	628.0000	10734.0000	6740952.00
18291	64	20	4	17	5	599.0000	14737.0000	8827463.00
18292	9	11	4	3	13	1201.0000	6874.0000	8255674.00
18293	4	1	4	20	12	604.0000	14153.0000	8548412.00
18294	66	15	4	20	11	941.0000	9224.0000	8679784.00
18295	82	5	4	4	16	1108.0000	11692.0000	12954736.00
18296	52	15	4	11	10	620.0000	13557.0000	8405340.00
18297	88	10	4	14	15	1036.0000	5396.0000	5590256.00
18298	50	11	4	13	15	961.0000	12025.0000	11556025.00
18299	66	6	4	8	10	919.0000	11717.0000	10767923.00
18300	2	5	4	11	6	1458.0000	13855.0000	20200590.00
18368	85	17	3	3	7	2.2330	20.0000	44.66
18369	87	7	1	7	3	3.3220	40.0000	132.88
18370	84	19	1	1	5	11.2200	10.0000	112.20
18371	86	9	5	5	1	33.1100	30.0000	993.30
18372	96	5	5	9	11	33.2100	18.0000	597.78
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Access source 1



After running the transformation

11 rows added



Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

InventoryKey	DateKey	ItemMasterKey	TransTypeKey	BranchPlantKey	CustVendorKey	UnitCost	Quantity	ExtCost
18300	2	5	4	11	6	1458.0000	13855.0000	20200590.00
18368	85	17	3	3	7	2.2330	20.0000	44.66
18369	87	7	1	7	3	3.3220	40.0000	132.88
18370	84	19	1	1	5	11.2200	10.0000	112.20
18371	86	9	5	5	1	33.1100	30.0000	993.30
18372	96	5	5	9	11	33.2100	18.0000	597.78
18373	14	14	4	10	18	76.3000	50.0000	3815.00
18374	11	10	2	16	12	29.0780	45.0000	1308.51
18375	6	4	4	2	8	21.7448	75.0000	1630.86
18376	7	6	4	6	14	24.1780	30.0000	725.34
18377	3	2	4	14	10	2.6890	55.0000	147.90
18378	13	12	2	18	2	1.8260	25.0000	45.65
18379	2	2	2	2	4	45.2000	10.0000	452.00
18380	8	6	4	8	10	25.3600	65.0000	1648.40
18381	1	1	2	12	12	24.3000	60.0000	1458.00
18382	5	4	2	16	6	14.4500	80.0000	1156.00
18383	16	20	2	18	8	22.0000	90.0000	1980.00
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Conclusion

During this project, I learnt how to extract the data from different types of sources (Excel and access sources), clean and filter the data based on conditions, using different step transformations available at the pentaho environment (excel / access input; filter rows; data validators; selected values; split fields) and then I was able to merge the data transformed into the data warehouse after matching the records with the values existing in the dimensions's tables and also calculating the measures using steps (merge join,sort table,calculator,table input) .

And now the data is ready for the final step "the data mining".