INFO 7370 Designing Data Architecture Business Intelligence

Assignment 4

Srishti Ashok Mishra

NU ID : 001305178

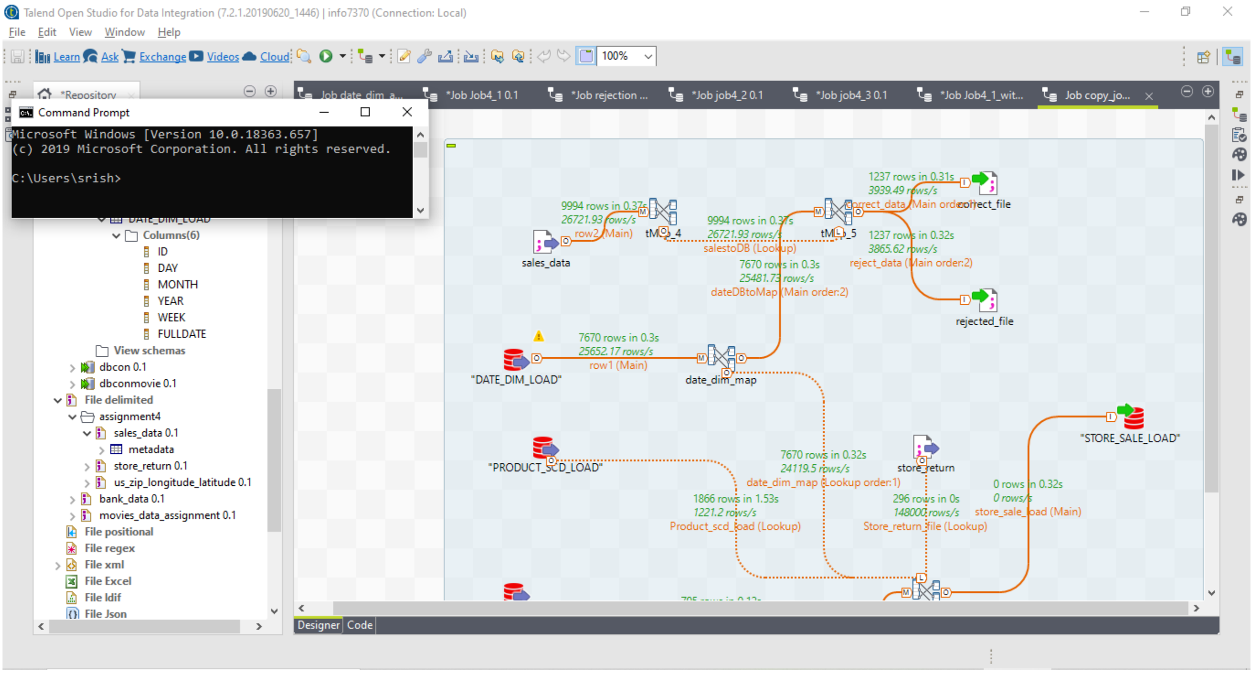
JOB:

SALES LOAD:

For this Job:

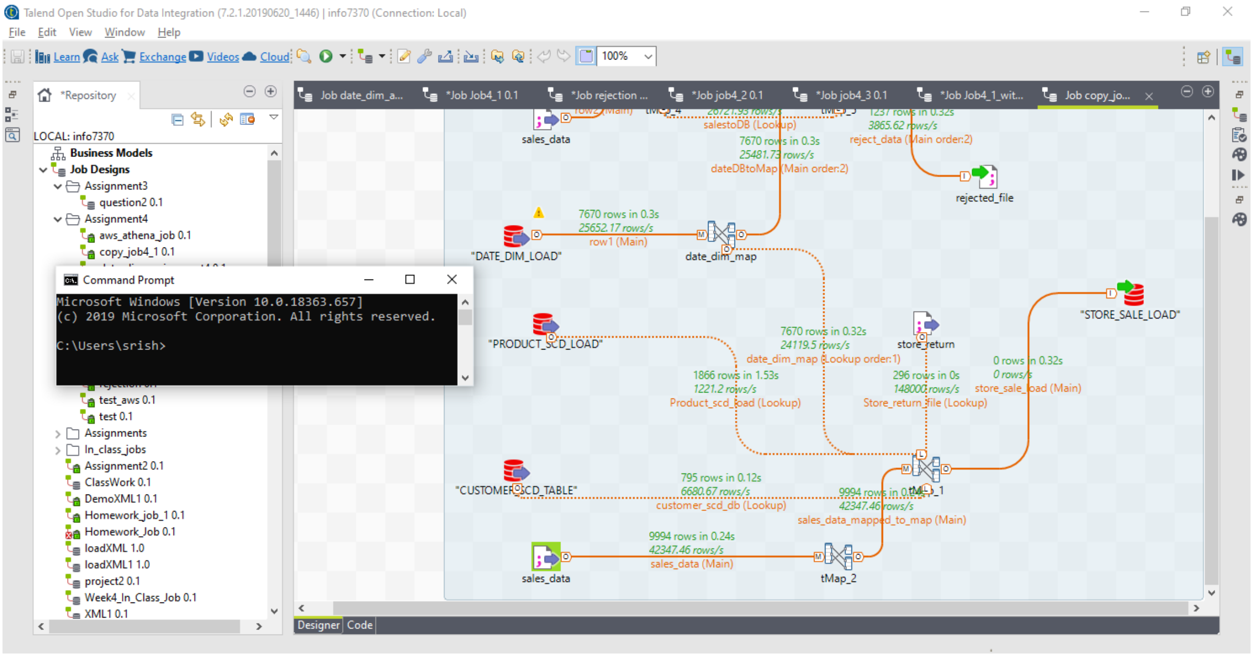
Customer_dim_id, *product_dim_id*, *order_dt_id* are based on the other jobs and are linked by surrogate key.

Ship_date, *ship_mode*, *qty_sold*, *unit_price*, *discount_percentage*, *order_id* where obtain and/ or calculated using store_sales file in tmap



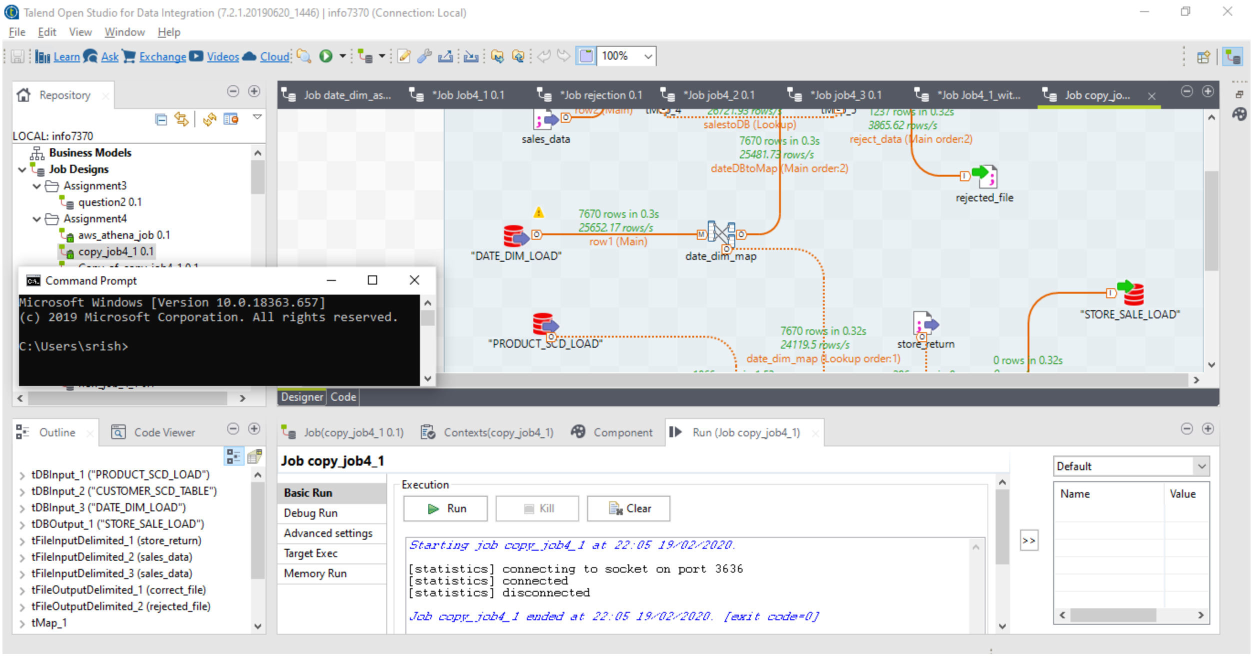
```
Microsoft Windows [Version 10.0.18363.657]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\srish>
```



```
Microsoft Windows [Version 10.0.18363.657]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\srish>
```



```
Microsoft Windows [Version 10.0.18363.657]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\srish>
```

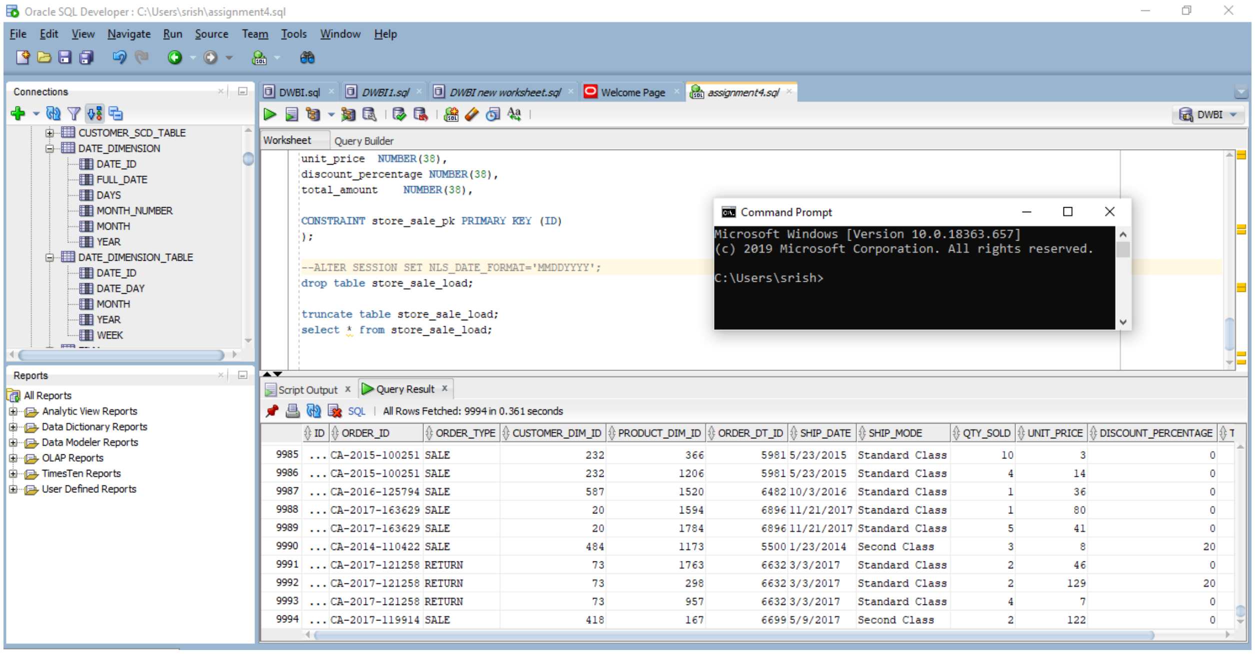
tDBInput_1 ("PRODUCT_SCD_LOAD")
tDBInput_2 ("CUSTOMER_SCD_TABLE")
tDBInput_3 ("DATE_DIM_LOAD")
tDBOutput_1 ("STORE_SALE_LOAD")
tFileInputDelimited_1 (store_return)
tFileInputDelimited_2 (sales_data)
tFileInputDelimited_3 (sales_data)
tFileOutputDelimited_1 (correct_file)
tFileOutputDelimited_2 (rejected_file)
tMap_1

```
Starting job copy_job4_1 at 22:05 19/02/2020.

[statistics] connecting to socket on port 3636
[statistics] connected
[statistics] disconnected

Job copy_job4_1 ended at 22:05 19/02/2020.  [exit code=0]
```

Data in DB:



Oracle SQL Developer : C:\Users\srish\assignment4.sql

```
unit_price  NUMBER(38),
discount_percentage NUMBER(38),
total_amount    NUMBER(38),

CONSTRAINT store_sale_pk PRIMARY KEY (ID)
);

--ALTER SESSION SET NLS_DATE_FORMAT='MMDDYYYY';
drop table store_sale_load;

truncate table store_sale_load;
select * from store_sale_load;
```

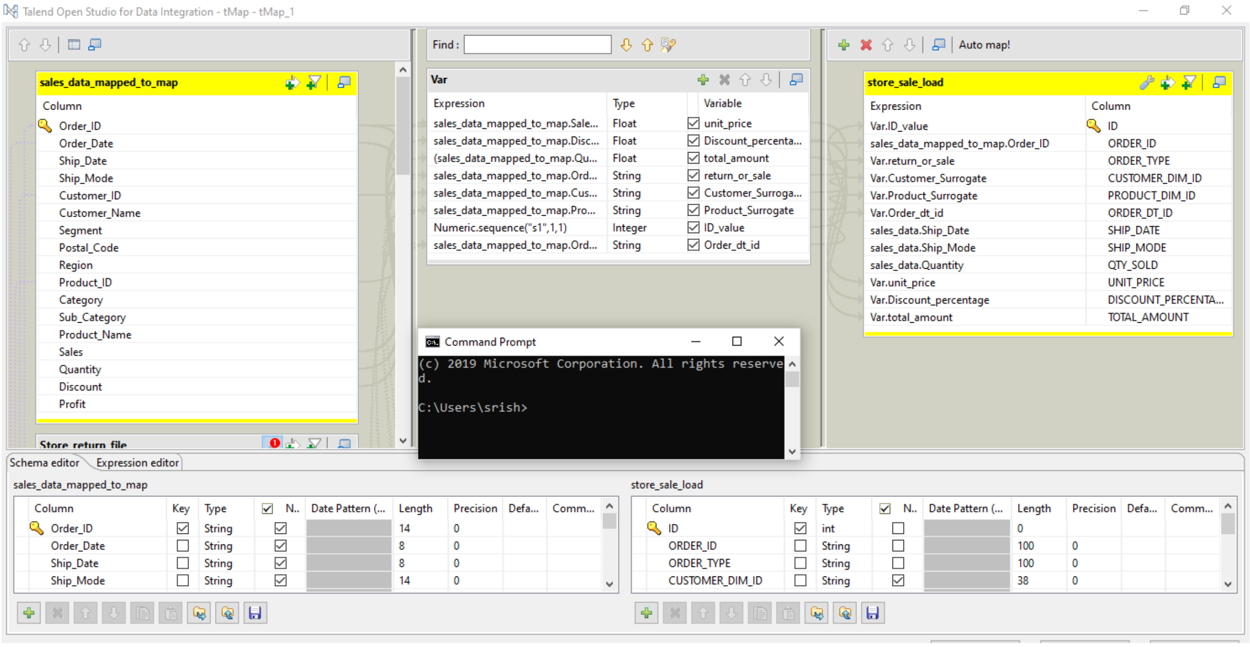
```
Microsoft Windows [Version 10.0.18363.657]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\srish>
```

All Rows Fetched: 9994 in 0.361 seconds

	ID	ORDER_ID	ORDER_TYPE	CUSTOMER_DIM_ID	PRODUCT_DIM_ID	ORDER_DT_ID	SHIP_DATE	SHIP_MODE	QTY_SOLD	UNIT_PRICE	DISCOUNT_PERCENTAGE
9985	...	CA-2015-100251	SALE	232	366	5981	5/23/2015	Standard Class	10	3	0
9986	...	CA-2015-100251	SALE	232	1206	5981	5/23/2015	Standard Class	4	14	0
9987	...	CA-2016-125794	SALE	587	1520	6482	10/3/2016	Standard Class	1	36	0
9988	...	CA-2017-163629	SALE	20	1594	6896	11/21/2017	Standard Class	1	80	0
9989	...	CA-2017-163629	SALE	20	1784	6896	11/21/2017	Standard Class	5	41	0
9990	...	CA-2014-110422	SALE	484	1173	5500	1/23/2014	Second Class	3	8	20
9991	...	CA-2017-121258	RETURN	73	1763	6632	3/3/2017	Standard Class	2	46	0
9992	...	CA-2017-121258	RETURN	73	298	6632	3/3/2017	Standard Class	2	129	20
9993	...	CA-2017-121258	RETURN	73	957	6632	3/3/2017	Standard Class	4	7	0
9994	...	CA-2017-119914	SALE	418	167	6699	5/9/2017	Second Class	2	122	0

MAPPING:



Talend Open Studio for Data Integration - tMap - tMap_1

sales_data_mapped_to_map

Column
Order_ID
Order_Date
Ship_Date
Ship_Mode
Customer_ID
Customer_Name
Segment
Postal_Code
Region
Product_ID
Category
Sub_Category
Product_Name
Sales
Quantity
Discount
Profit

Var

Expression	Type	Variable
sales_data_mapped_to_map.Sale...	Float	☑ unit_price
sales_data_mapped_to_map.Disc...	Float	☑ Discount_percenta...
(sales_data_mapped_to_map.Qu...	Float	☑ total_amount
sales_data_mapped_to_map.Ord...	String	☑ return_or_sale
sales_data_mapped_to_map.Cus...	String	☑ Customer_Surroga...
sales_data_mapped_to_map.Pro...	String	☑ Product_Surrogate
Numeric.sequence("s1",1,1)	Integer	☑ ID_value
sales_data_mapped_to_map.Ord...	String	☑ Order_dt_id

store_sale_load

Expression	Column
Var.ID_value	ID
sales_data_mapped_to_map.Order_ID	ORDER_ID
Var.return_or_sale	ORDER_TYPE
Var.Customer_Surrogate	CUSTOMER_DIM_ID
Var.Product_Surrogate	PRODUCT_DIM_ID
Var.Order_dt_id	ORDER_DT_ID
sales_data.Ship_Date	SHIP_DATE
sales_data.Ship_Mode	SHIP_MODE
sales_data.Quantity	QTY_SOLD
Var.unit_price	UNIT_PRICE
Var.Discount_percentage	DISCOUNT_PERCENTA...
Var.total_amount	TOTAL_AMOUNT

Store_return_file

```
(c) 2019 Microsoft Corporation. All rights reserve
d.

C:\Users\srish>
```

Schema editor / Expression editor

sales_data_mapped_to_map

Column	Key	Type	☑ N..	Date Pattern (...	Length	Precision	Defa...	Comm...
Order_ID	☑	String	☑		14	0		
Order_Date	☐	String	☑		8	0		
Ship_Date	☐	String	☑		8	0		
Ship_Mode	☐	String	☑		14	0		

store_sale_load

Column	Key	Type	☑ N..	Date Pattern (...	Length	Precision	Defa...	Comm...
ID	☑	int	☐		0			
ORDER_ID	☐	String	☐		100	0		
ORDER_TYPE	☐	String	☐		100	0		
CUSTOMER_DIM_ID	☐	String	☑		38	0		

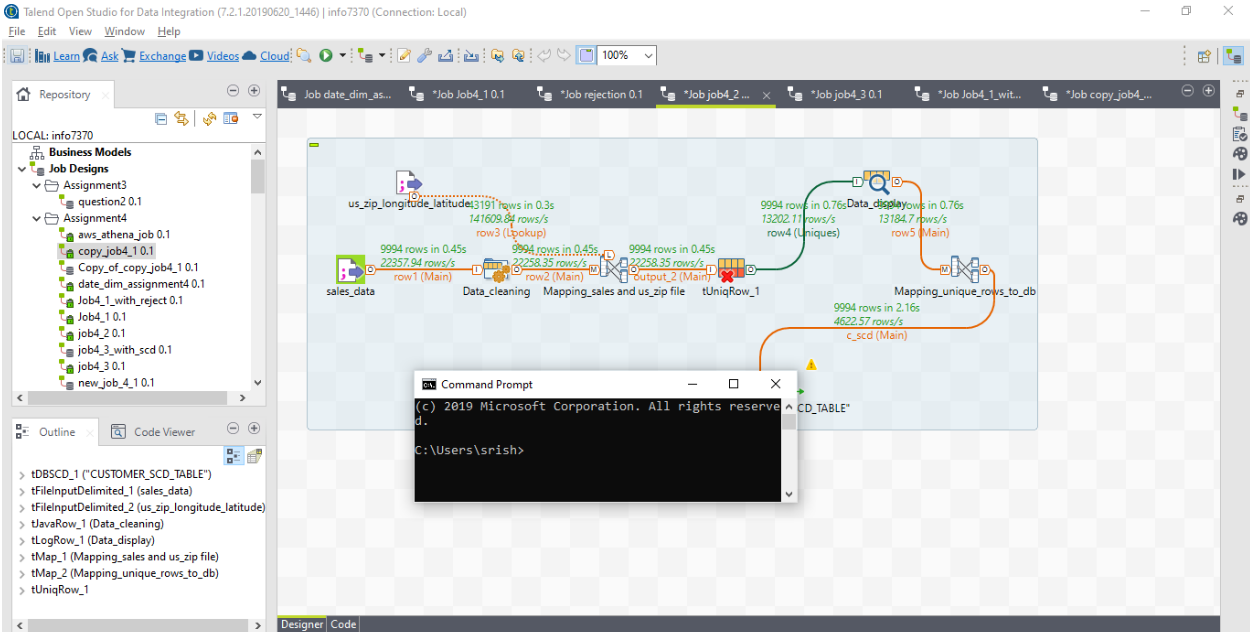
CUSTOMER_LOAD:

For this Job:

Data for *latitude*, *Longitude*, *city*, *state* was obtained by from file us-zip-code-latitude-and-longitude.csv lookup based on zipcode

Data for *First name* and *last name* was obtained from splitting up full name in *tmap* component

Data for *segment* and *region* was directly from *store_sales* file by using it as a delimited component.



Talend Open Studio for Data Integration (7.2.1.20190620_1446) | info7370 (Connection: Local)

LOCAL: info7370

- Business Models
- Job Designs
  - Assignment3
    - question2 0.1
  - Assignment4
    - aws_athena_job 0.1
    - copy_job4_1 0.1
    - Copy_of_copy_job4_1 0.1
    - date_dim_assignment4 0.1
    - Job4_1_with_reject 0.1
    - Job4_1 0.1
    - job4_2 0.1
    - job4_3_with_scd 0.1
    - job4_3 0.1
    - new_job_4_1 0.1

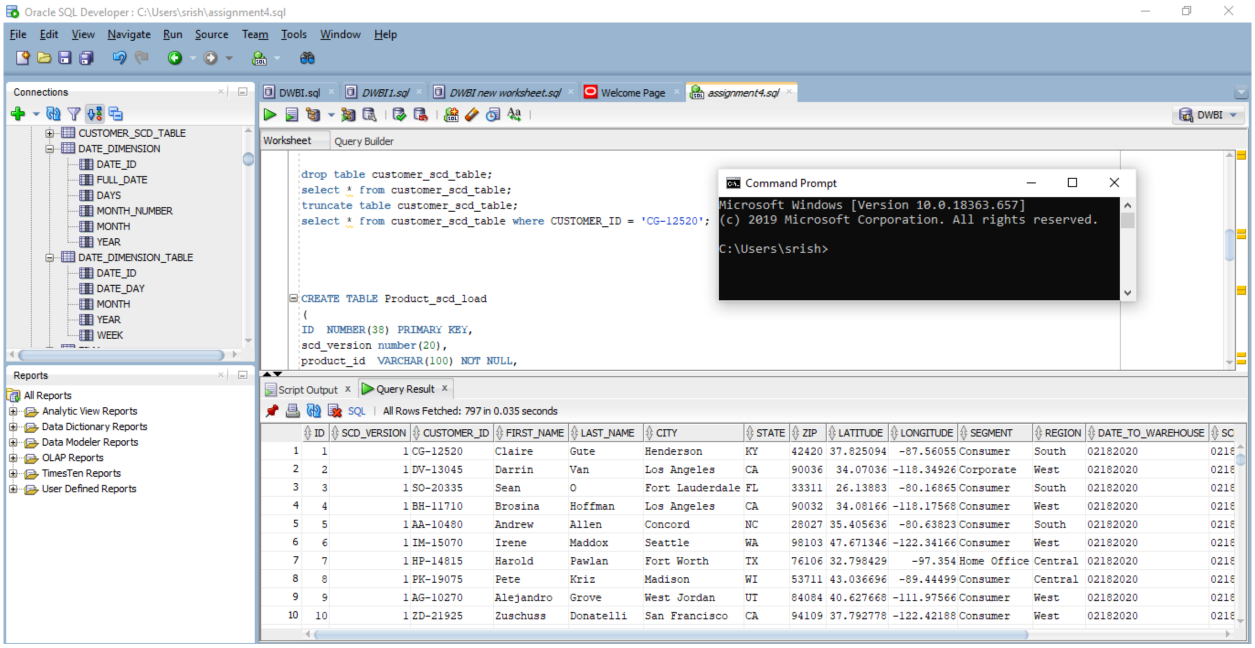
- tDBSCD_1 ("CUSTOMER_SCD_TABLE")
- tFileInputDelimited_1 (sales_data)
- tFileInputDelimited_2 (us_zip_longitude_latitude)
- tJavaRow_1 (Data_cleaning)
- tLogRow_1 (Data_display)
- tMap_1 (Mapping_sales and us_zip file)
- tMap_2 (Mapping_unique_rows_to_db)
- tUniqRow_1

Command Prompt

```
(c) 2019 Microsoft Corporation. All rights reserve
d.

C:\Users\srish>
```

Data in DB:



```
drop table customer_scd_table;
select * from customer_scd_table;
truncate table customer_scd_table;
select * from customer_scd_table where CUSTOMER_ID = 'CG-12520';



CREATE TABLE Product_scd_load
(
ID  NUMBER(38) PRIMARY KEY,
scd_version number(20),
product_id  VARCHAR(100) NOT NULL,
```

Command Prompt

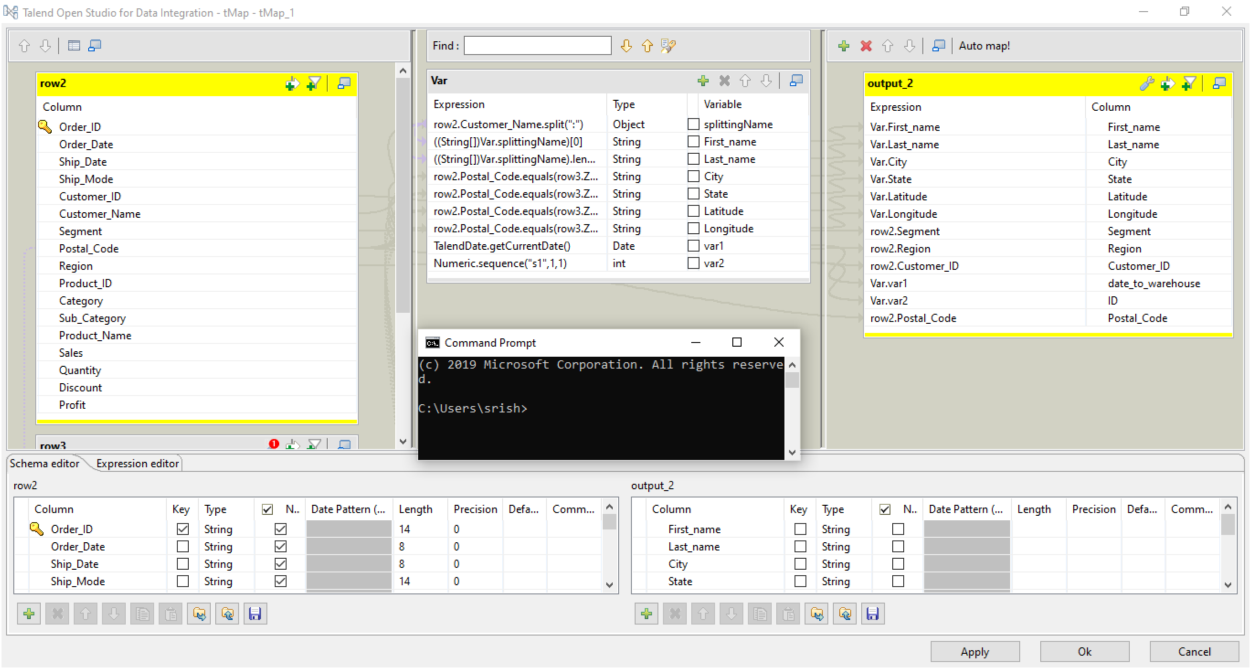
```
Microsoft Windows [Version 10.0.18363.657]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\srish>
```

All Rows Fetched: 797 in 0.035 seconds

	ID	SCD_VERSION	CUSTOMER_ID	FIRST_NAME	LAST_NAME	CITY	STATE	ZIP	LATITUDE	LONGITUDE	SEGMENT	REGION	DATE_TO_WAREHOUSE	SC
1	1	1	CG-12520	Claire	Gute	Henderson	KY	42420	37.825094	-87.56055	Consumer	South	02182020	0218
2	2	1	DV-13045	Darrin	Van	Los Angeles	CA	90036	34.07036	-118.34926	Corporate	West	02182020	0218
3	3	1	SO-20335	Sean	O	Fort Lauderdale	FL	33311	26.13883	-80.16865	Consumer	South	02182020	0218
4	4	1	BH-11710	Brosina	Hoffman	Los Angeles	CA	90032	34.08166	-118.17568	Consumer	West	02182020	0218
5	5	1	AA-10480	Andrew	Allen	Concord	NC	28027	35.405636	-80.63823	Consumer	South	02182020	0218
6	6	1	IM-15070	Irene	Maddox	Seattle	WA	98103	47.671346	-122.34166	Consumer	West	02182020	0218
7	7	1	HP-14815	Harold	Pawlan	Fort Worth	TX	76106	32.798429	-97.354	Home Office	Central	02182020	0218
8	8	1	PK-19075	Pete	Kriz	Madison	WI	53711	43.036696	-89.44499	Consumer	Central	02182020	0218
9	9	1	AG-10270	Alejandro	Grove	West Jordan	UT	84084	40.627668	-111.97566	Consumer	West	02182020	0218
10	10	1	ZD-21925	Zuschuss	Donatelli	San Francisco	CA	94109	37.792778	-122.42188	Consumer	West	02182020	0218

MAPPING:



row2

Column
Order_ID
Order_Date
Ship_Date
Ship_Mode
Customer_ID
Customer_Name
Segment
Postal_Code
Region
Product_ID
Category
Sub_Category
Product_Name
Sales
Quantity
Discount
Profit

Var

Expression	Type		Variable
row2.Customer_Name.split(":")	Object	□	splittingName
((String[])Var.splittingName)[0]	String	□	First_name
((String[])Var.splittingName).len...	String	□	Last_name
row2.Postal_Code.equals(row3.Z...	String	□	City
row2.Postal_Code.equals(row3.Z...	String	□	State
row2.Postal_Code.equals(row3.Z...	String	□	Latitude
row2.Postal_Code.equals(row3.Z...	String	□	Longitude
TalendDate.getCurrentDate()	Date	□	var1
Numeric.sequence("s1",1,1)	int	□	var2

output_2

Expression	Column
Var.First_name	First_name
Var.Last_name	Last_name
Var.City	City
Var.State	State
Var.Latitude	Latitude
Var.Longitude	Longitude
row2.Segment	Segment
row2.Region	Region
row2.Customer_ID	Customer_ID
Var.var1	date_to_warehouse
Var.var2	ID
row2.Postal_Code	Postal_Code

Command Prompt

```
(c) 2019 Microsoft Corporation. All rights reserve
d.

C:\Users\srish>
```

row2

Column	Key	Type	☑ N..	Date Pattern (...	Length	Precision	Defa...	Comm...
Order_ID	☑	String	☑		14	0		
Order_Date	□	String	☑		8	0		
Ship_Date	□	String	☑		8	0		
Ship_Mode	□	String	☑		14	0		

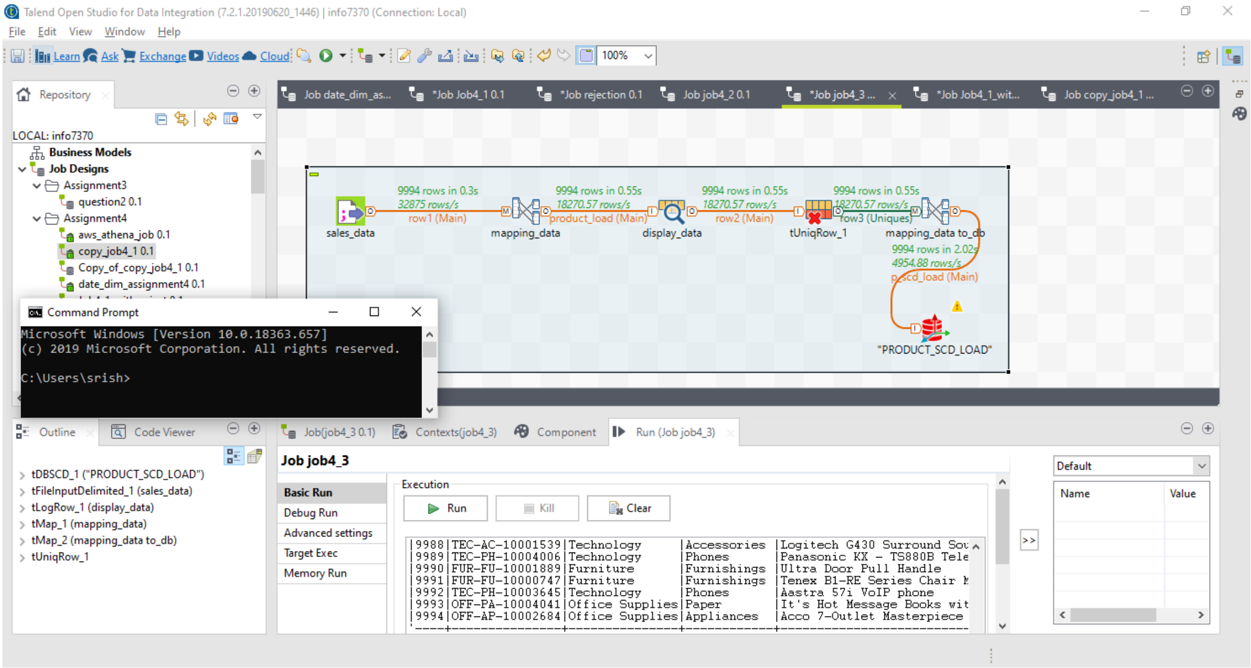
output_2

Column	Key	Type	☑ N..	Date Pattern (...	Length	Precision	Defa...	Comm...
First_name	□	String	□					
Last_name	□	String	□					
City	□	String	□					
State	□	String	□					

PRODUCT_LOAD:

For this Job:

All the data was directly from *store_sales* file by using it as a delimited component.



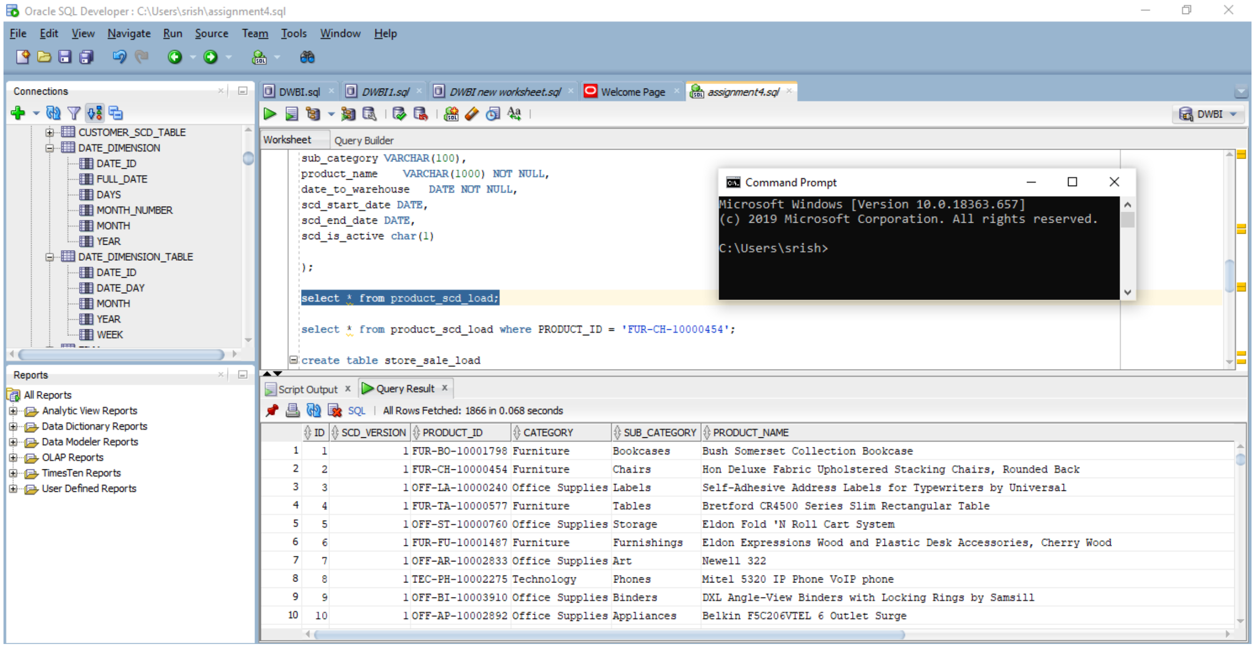
```
Microsoft Windows [Version 10.0.18363.657]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\srish>
```

```
tDBSCD_1 ("PRODUCT_SCD_LOAD")
tFileInputDelimited_1 (sales_data)
tLogRow_1 (display_data)
tMap_1 (mapping_data)
tMap_2 (mapping_data to_db)
tUniqRow_1
```

```
|9988|TEC-AC-10001539|Technology     |Accessories |Logitech G430 Surround Sou
|9989|TEC-PH-10004006|Technology     |Phones      |Panasonic KX - TS880B Tele
|9990|FUR-FU-10001889|Furniture      |Furnishings |Ultra Door Pull Handle
|9991|FUR-FU-10000747|Furniture      |Furnishings |Tenex B1-RE Series Chair M
|9992|TEC-PH-10003645|Technology     |Phones      |Aastra 57i VoIP phone
|9993|OFF-PA-10004041|Office Supplies|Paper       |It's Hot Message Books wit
|9994|OFF-AP-10002684|Office Supplies|Appliances  |Acco 7-Outlet Masterpiece
'----+---------------+---------------+------------+--------------------------
```

Data in DB:



```
sub_category VARCHAR(100),
product_name    VARCHAR(1000) NOT NULL,
date_to_warehouse   DATE NOT NULL,
scd_start_date DATE,
scd_end_date DATE,
scd_is_active char(1)

);

select * from product_scd_load;

select * from product_scd_load where PRODUCT_ID = 'FUR-CH-10000454';

create table store_sale_load
```

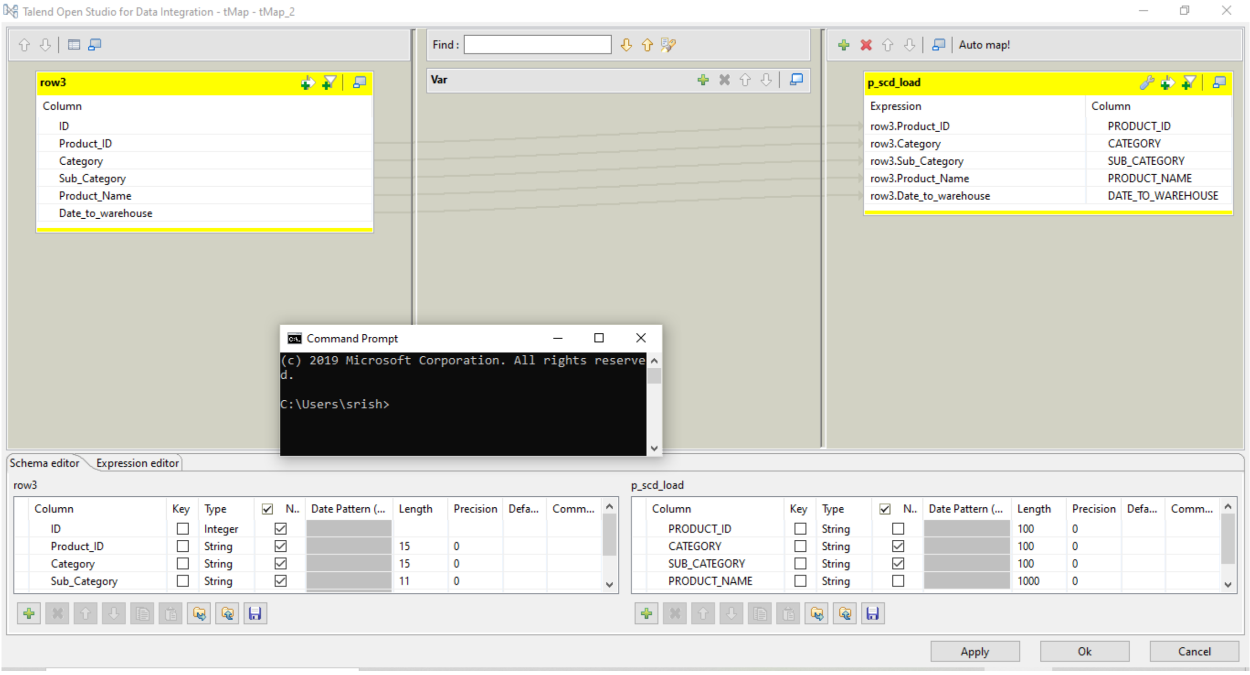
```
Microsoft Windows [Version 10.0.18363.657]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\srish>
```

All Rows Fetched: 1866 in 0.068 seconds

	ID	SCD_VERSION	PRODUCT_ID	CATEGORY	SUB_CATEGORY	PRODUCT_NAME
1	1	1	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase
2	2	1	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back
3	3	1	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters by Universal
4	4	1	FUR-TA-10000577	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table
5	5	1	OFF-ST-10000760	Office Supplies	Storage	Eldon Fold 'N Roll Cart System
6	6	1	FUR-FU-10001487	Furniture	Furnishings	Eldon Expressions Wood and Plastic Desk Accessories, Cherry Wood
7	7	1	OFF-AR-10002833	Office Supplies	Art	Newell 322
8	8	1	TEC-PH-10002275	Technology	Phones	Mitel 5320 IP Phone VoIP phone
9	9	1	OFF-BI-10003910	Office Supplies	Binders	DXL Angle-View Binders with Locking Rings by Samsill
10	10	1	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C206VTEL 6 Outlet Surge

MAPPING:



row3
Column
ID
Product_ID
Category
Sub_Category
Product_Name
Date_to_warehouse

p_scd_load

Expression	Column
row3.Product_ID	PRODUCT_ID
row3.Category	CATEGORY
row3.Sub_Category	SUB_CATEGORY
row3.Product_Name	PRODUCT_NAME
row3.Date_to_warehouse	DATE_TO_WAREHOUSE

```
(c) 2019 Microsoft Corporation. All rights reserve
d.

C:\Users\srish>
```

row3

Column	Key	Type	☑ N..	Date Pattern (...	Length	Precision	Defa...	Comm...
ID	□	Integer	☑					
Product_ID	□	String	☑		15	0		
Category	□	String	☑		15	0		
Sub_Category	□	String	☑		11	0		

p_scd_load

Column	Key	Type	☑ N..	Date Pattern (...	Length	Precision	Defa...	Comm...
PRODUCT_ID	□	String	□		100	0		
CATEGORY	□	String	☑		100	0		
SUB_CATEGORY	□	String	☑		100	0		
PRODUCT_NAME	□	String	□		1000	0		

SCD EDITOR:

Here the implementation is based on SCD-2.

Product_load

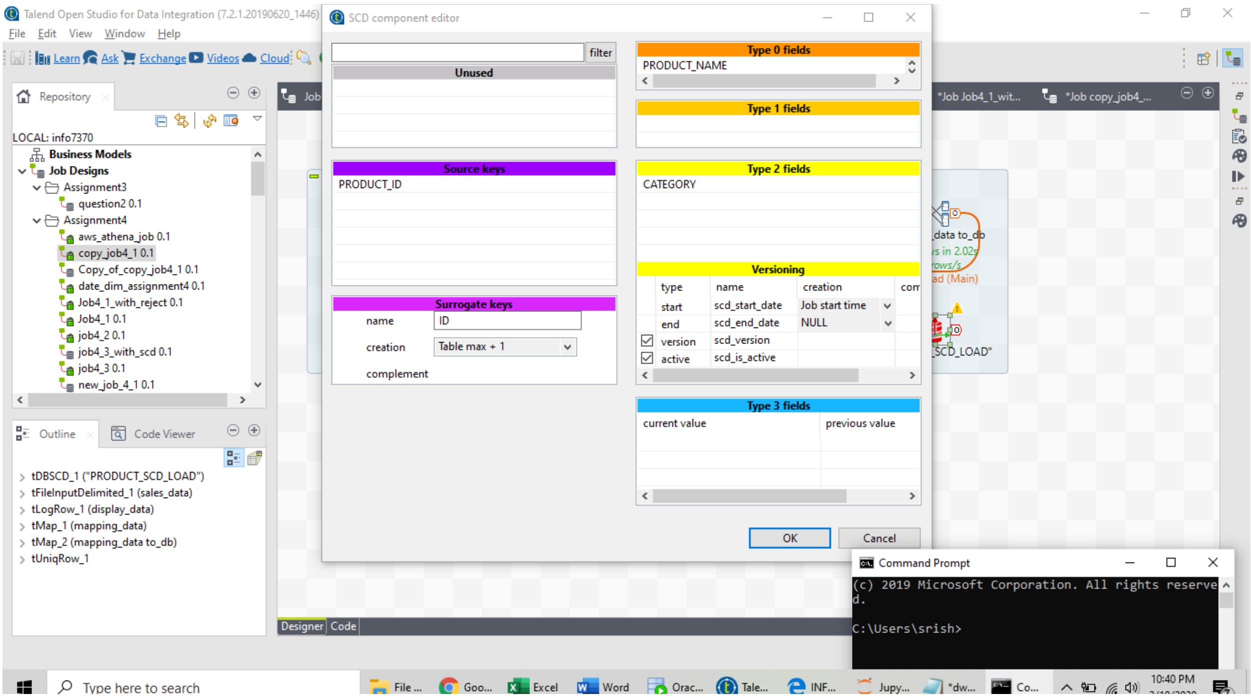
The source delimited file contains various details including *product_id*, *category*, *sub_category*, *product_name*. An *id* column helps ensuring the uniqueness of the data.

Here source Id is *product_id*

Scd type 2 field is *category*

There is a new record inserted in the dimensional table with a separate key each time the category changes.

The Table will retain both old and new value and the latest value will be shown active in the *scd_is_active* column.



Talend Open Studio for Data Integration (7.2.1.20190620_1446)

SCD component editor

Repository — LOCAL: info7370
- Business Models
- Job Designs
  - Assignment3
    - question2 0.1
  - Assignment4
    - aws_athena_job 0.1
    - copy_job4_1 0.1
    - Copy_of_copy_job4_1 0.1
    - date_dim_assignment4 0.1
    - Job4_1_with_reject 0.1
    - Job4_1 0.1
    - job4_2 0.1
    - job4_3_with_scd 0.1
    - job4_3 0.1
    - new_job_4_1 0.1

Outline
- tDBSCD_1 ("PRODUCT_SCD_LOAD")
- tFileInputDelimited_1 (sales_data)
- tLogRow_1 (display_data)
- tMap_1 (mapping_data)
- tMap_2 (mapping_data to_db)
- tUniqRow_1

Unused:

Source keys: PRODUCT_ID

Surrogate keys: name: ID; creation: Table max + 1; complement:

Type 0 fields: PRODUCT_NAME

Type 1 fields:

Type 2 fields: CATEGORY

Versioning

	type	name	creation
	start	scd_start_date	Job start time
	end	scd_end_date	NULL
☑	version	scd_version	
☑	active	scd_is_active	

Type 3 fields: current value | previous value

OK | Cancel

```
Command Prompt
(c) 2019 Microsoft Corporation. All rights reserve
d.

C:\Users\srish>
```

Here the SCD type 2 columns are:

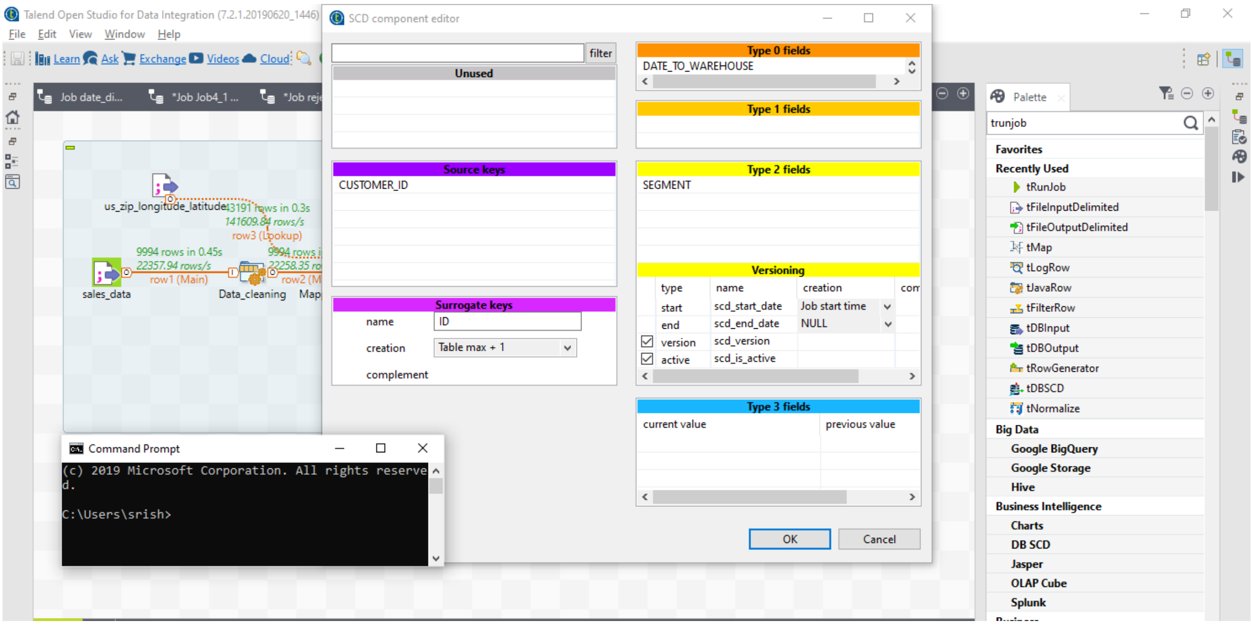
Scd_start_date to store the start date

SCD_end_date to store the end date

SCD_is_active stores which of the column is active

SCD_version to store the version of the scd

Similarly for Customer_load



Talend Open Studio for Data Integration (7.2.1.20190620_1446)

SCD component editor

Unused:

Source keys: CUSTOMER_ID

Surrogate keys: name: ID; creation: Table max + 1; complement:

Type 0 fields: DATE_TO_WAREHOUSE

Type 1 fields:

Type 2 fields: SEGMENT

Versioning

	type	name	creation
	start	scd_start_date	Job start time
	end	scd_end_date	NULL
☑	version	scd_version	
☑	active	scd_is_active	

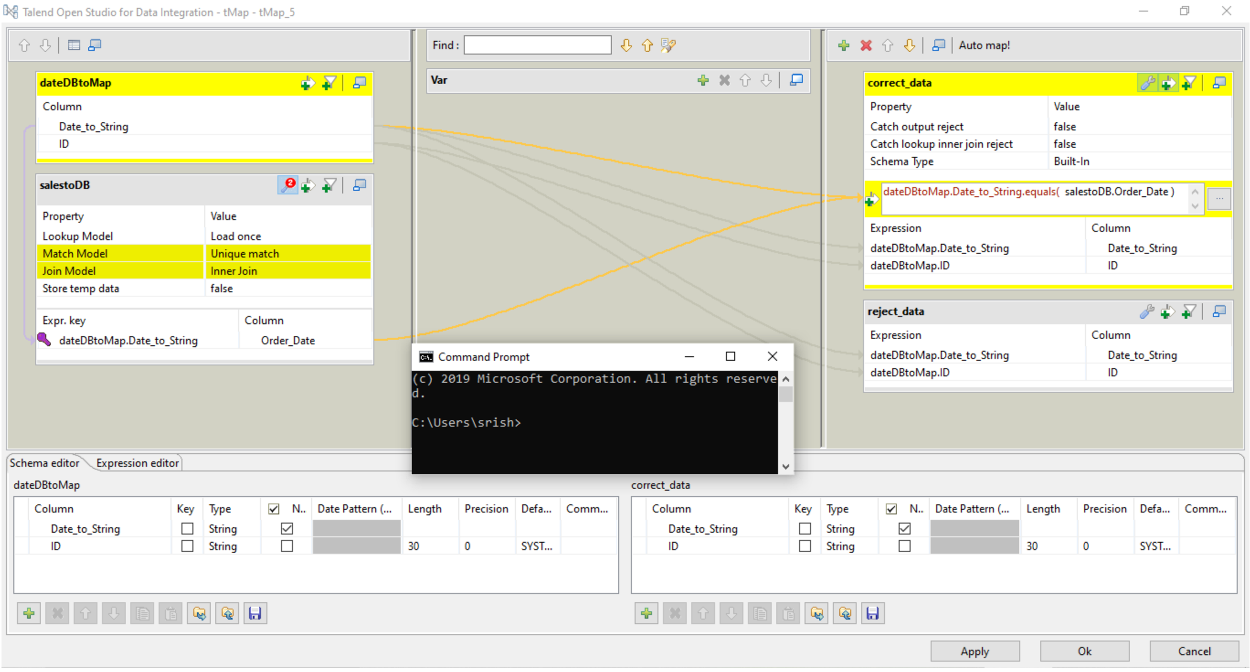
Type 3 fields: current value | previous value

OK | Cancel

```
Command Prompt
(c) 2019 Microsoft Corporation. All rights reserve
d.

C:\Users\srish>
```

REJECTED DATA MAPPING:



Talend Open Studio for Data Integration - tMap - tMap_5

dateDBtoMap

Column
Date_to_String
ID

salestoDB

Property	Value
Lookup Model	Load once
Match Model	Unique match
Join Model	Inner Join
Store temp data	false

Expr. key	Column
dateDBtoMap.Date_to_String	Order_Date

correct_data

Property	Value
Catch output reject	false
Catch lookup inner join reject	false
Schema Type	Built-In

Filter: dateDBtoMap.Date_to_String.equals(salestoDB.Order_Date)

Expression	Column
dateDBtoMap.Date_to_String	Date_to_String
dateDBtoMap.ID	ID

reject_data

Expression	Column
dateDBtoMap.Date_to_String	Date_to_String
dateDBtoMap.ID	ID

Command Prompt

```
(c) 2019 Microsoft Corporation. All rights reserve
d.

C:\Users\srish>
```

Schema editor | Expression editor

dateDBtoMap

Column	Key	Type	☑ N..	Date Pattern (...	Length	Precision	Defa...	Comm...
Date_to_String	□	String	☑					
ID	□	String	□		30	0	SYST...	

correct_data

Column	Key	Type	☑ N..	Date Pattern (...	Length	Precision	Defa...	Comm...
Date_to_String	□	String	☑					
ID	□	String	□		30	0	SYST...	

Here the file is rejected based on the order_id and date_dimension table.

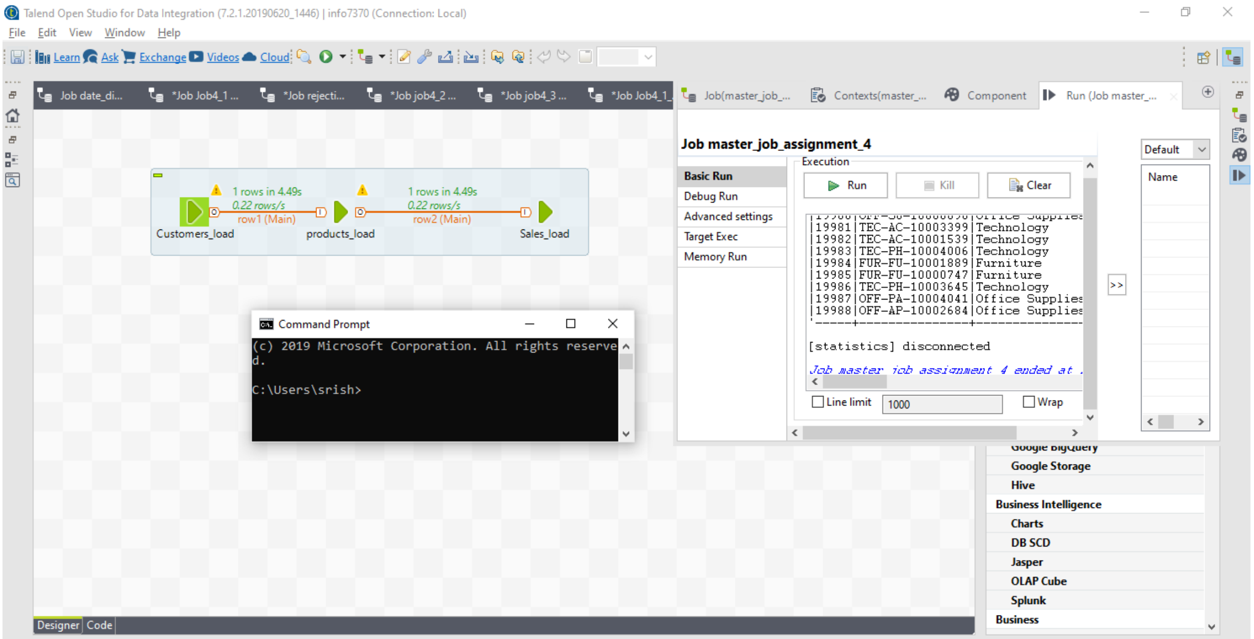
dateDBtoMap.Date_to_String.equals(salestoDB.Order_Date)

if the data falls into the formula, the value will be a part of the correct data, else goes into the reject data csv.

Master Job:

Master Job consist of Customers_load followed by products_load job. Once these both are successfully completed then it invoke sales_load

It uses tRunJob component to merge all the jobs into a master job



Talend Open Studio for Data Integration (7.2.1.20190620_1446) | info7370 (Connection: Local)

Customers_load → products_load → Sales_load

1 rows in 4.49s, 0.22 rows/s, row1 (Main)

1 rows in 4.49s, 0.22 rows/s, row2 (Main)

Command Prompt

```
(c) 2019 Microsoft Corporation. All rights reserve
d.

C:\Users\srish>
```

Job master_job_assignment_4

Execution

```
|19981|TEC-AC-10003399|Technology
|19982|TEC-AC-10001539|Technology
|19983|TEC-PH-10004006|Technology
|19984|FUR-FU-10001889|Furniture
|19985|FUR-FU-10000747|Furniture
|19986|TEC-PH-10003645|Technology
|19987|OFF-PA-10004041|Office Supplies
|19988|OFF-AP-10002684|Office Supplies
'-----+---------------+---------------

[statistics] disconnected

Job master job assignment 4 ended at .
```

□ Line limit 1000 □ Wrap