Data Warehousing and Business Intelligence

IT3021

Assignment 1

Delivery Center: Food & Goods orders in Brazil

2022
IT20220860
Karawita K.S.A

Contents

Data Set Selection

https://www.kaggle.com/datasets/nosbielcs/brazilian-delivery-center

The Delivery Center is a platform that integrates retailers and marketplaces, creating a healthy ecosystem for sales of goods and food in Brazilian retail. We currently have a register with more than 14000 items. Thousands of orders and deliveries are processed daily with a network of thousands of merchants and delivery partners spread across all country regions.

All this generates data and more data all the time. In view of this, this network business is increasingly data-driven, that is, using data to make decisions and in a vision of the future, we know that using data intelligently can be our great differential in the market.

Channels: This dataset has information about the sales channels (marketplaces) where our retailers' goods and food are sold.

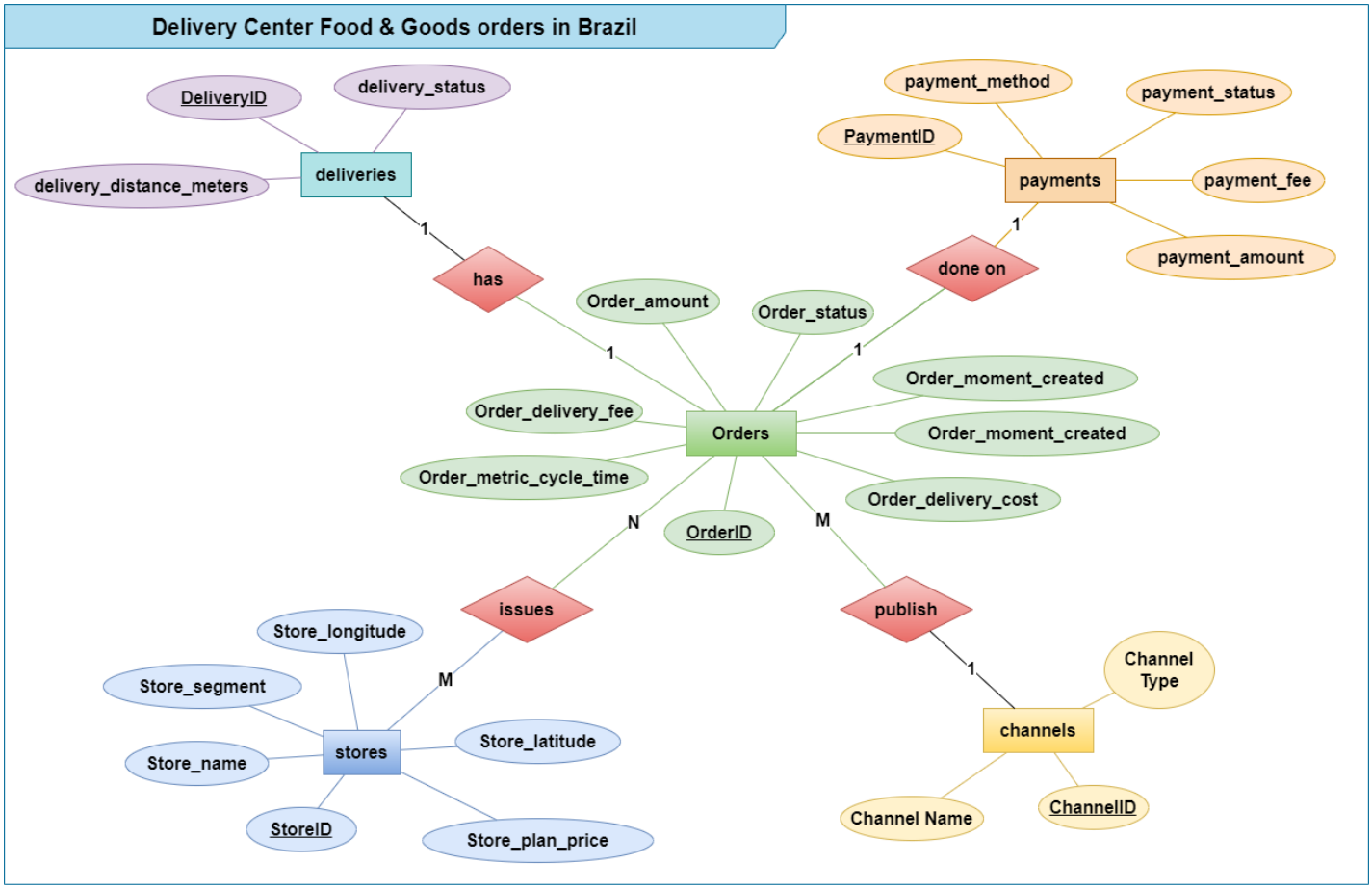
Deliveries: This dataset has information about deliveries made by our partner delivery partners.

Orders: This dataset has information about sales processed through the Delivery Center platform.

Payments: This dataset has information about payments made to the Delivery Center.

Stores: This dataset has information about the store owners. They use the Delivery Center Platform to sell their items on marketplaces.

ER diagram

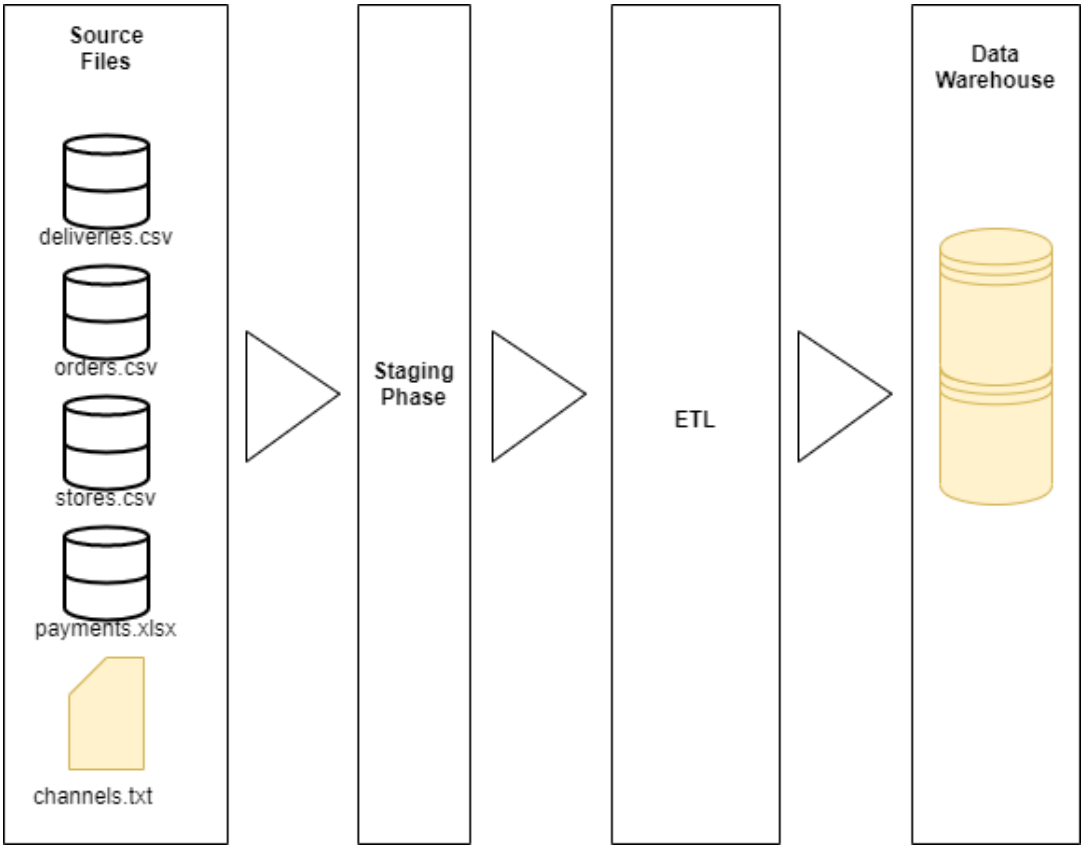

Delivery Center Food & Goods orders in Brazil
DeliveryID
delivery_status
delivery_distance_meters
deliveries
1
has
1
payment_method
payment_status
PaymentID
payments
payment_fee
payment_amount
1
done on
1
Order_amount
Order_status
Order_moment_created
Order_delivery_fee
Orders
Order_moment_created
Order_metric_cycle_time
Order_delivery_cost
N
OrderID
M
issues
publish
Store_longitude
Store_segment
M
1
Channel Type
stores
Store_latitude
channels
Store_name
Channel Name
ChannelID
StoreID
Store_plan_price

Preparation of Data Sources

Name	Date modified	Type	Size
channels.txt	5/16/2022 9:52 PM	Text Document	2 KB
deliveries.csv	5/17/2022 11:59 AM	Microsoft Excel C...	349 KB
orders.csv	5/21/2022 3:17 AM	Microsoft Excel C...	1,382 KB
payments.xlsx	5/17/2022 11:59 AM	Microsoft Excel W...	440 KB
stores.csv	5/16/2022 9:42 PM	Microsoft Excel C...	45 KB

The original data set contained more than 321 000 in some files. The database was varied with various types of data source files including csv, excel, txt. I filtered data to contain not more than 15000 records as it was more than enough to develop the solution perfectly.

Solution Architecture



Data sources containing data of orders, stores, and deliveries are in csv formats. They were imported as flat files to the source DB. Then in the staging phase the excel file and the text file containing payment data and channels data respectively were inserted in to the staging DB. And finally to the Data Warehouse through the ETL process.

Design and Development

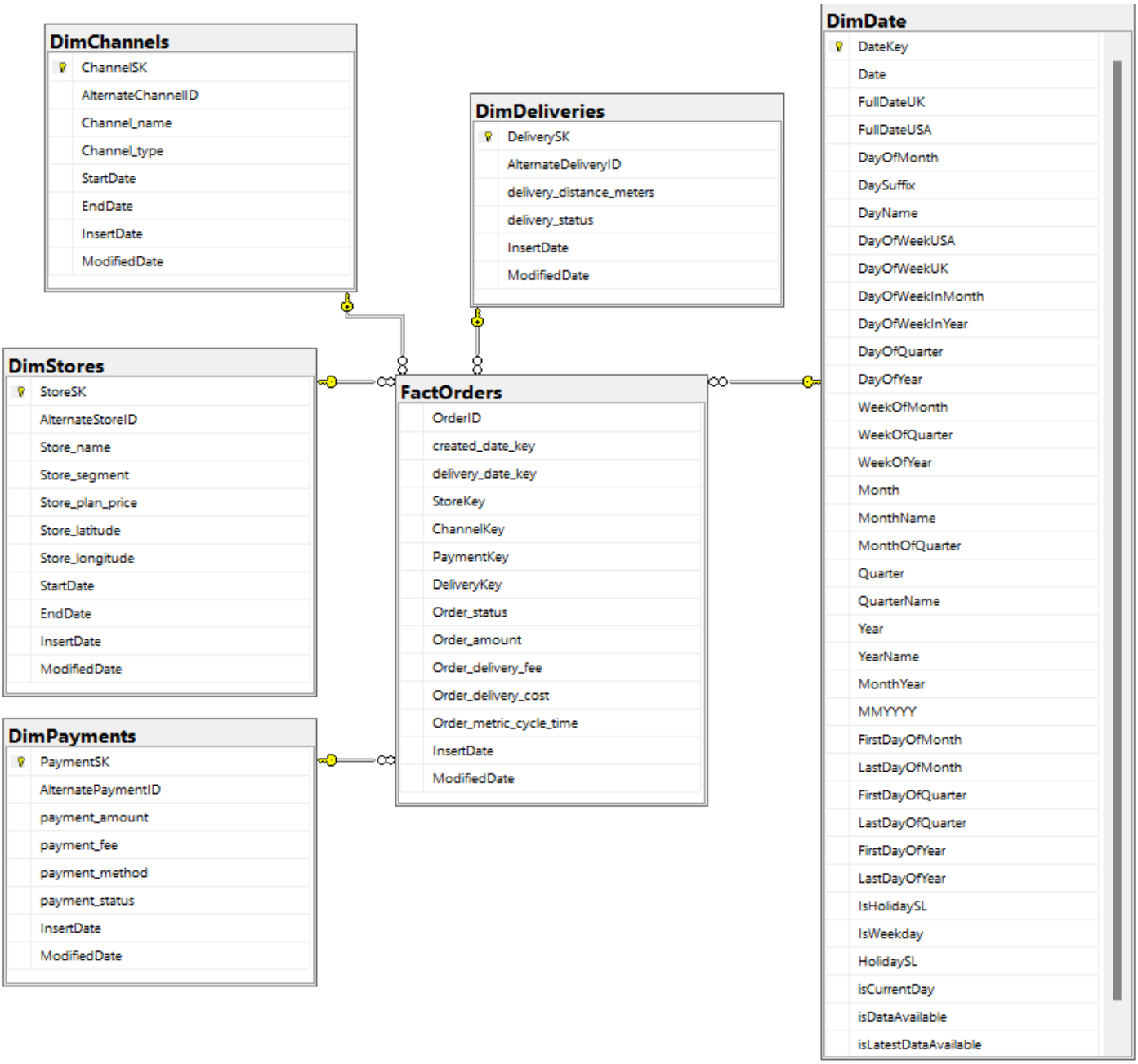
Dimensional Model

Star schema was selected to design the Data Warehouse for Delivery Center Food & Goods orders in Brazil. There are mainly four, dimensional tables and the fact table. All these dimensional tables are linked with the fact table.

Dimension tables and fact table:

- DimChannels
- DimDeliveries
- DimPayments
- DimStores
- FactOrders

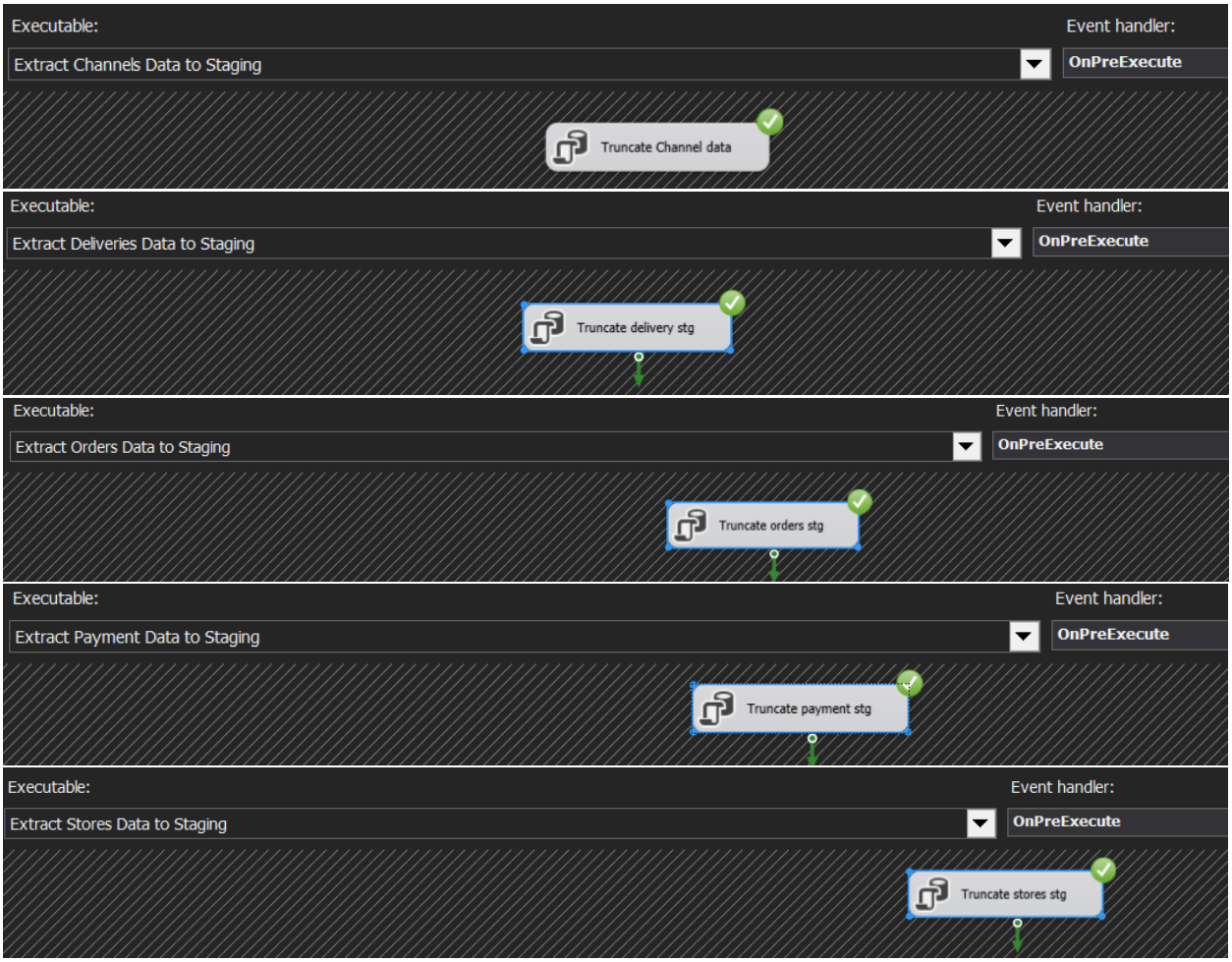
Hierarchies



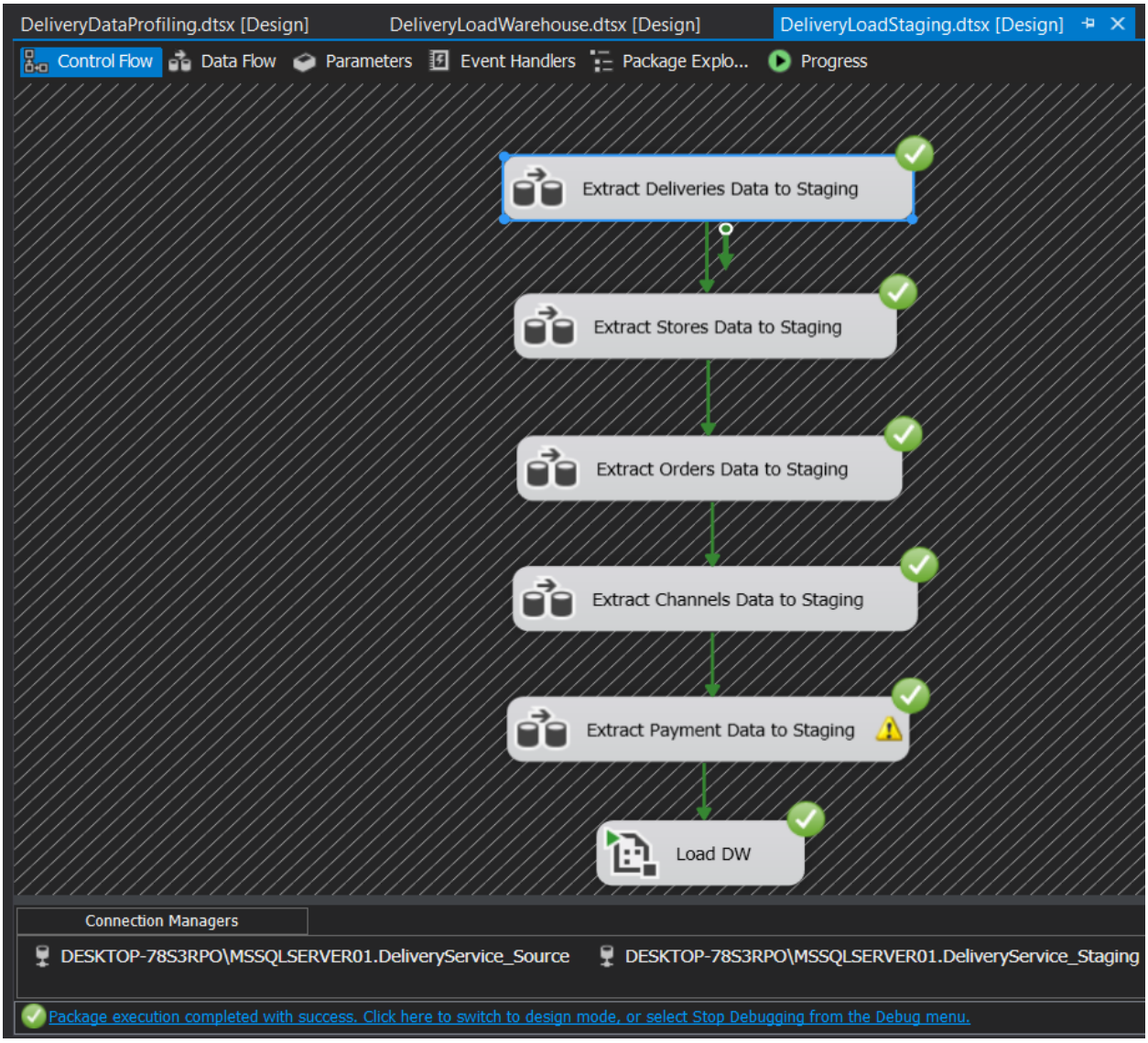
ETL Development Process

Truncating Staging tables

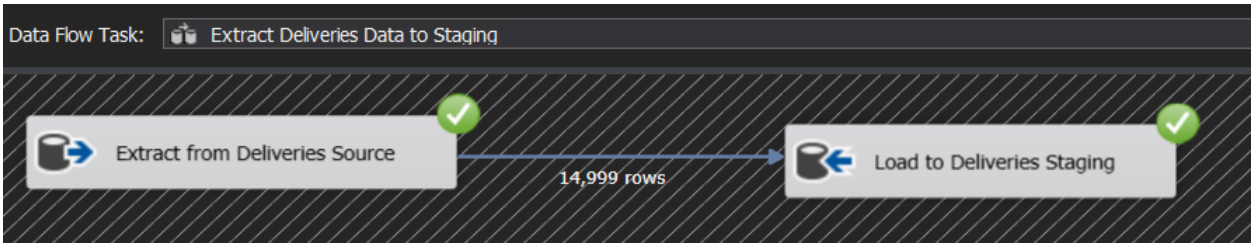
It is required to truncate processes for loading Staging tables with PreExecute Event Hander. That deletes all the existing rows in the target table before loading any new data. This process prevents data duplicating



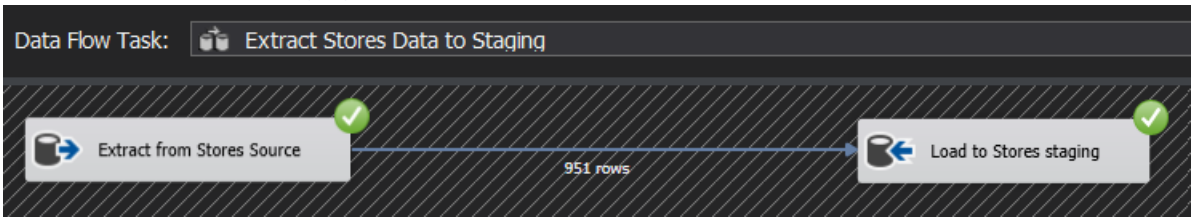
Loading to Staging



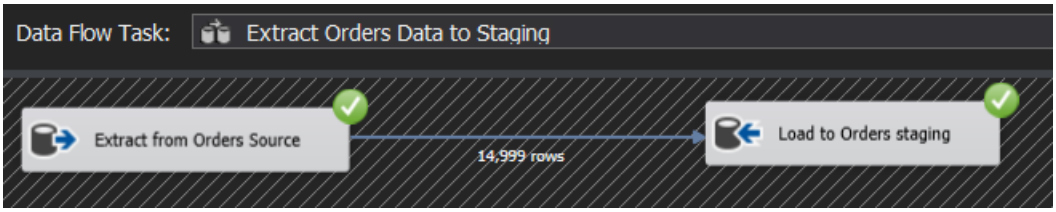
Extract Deliveries Data to Staging



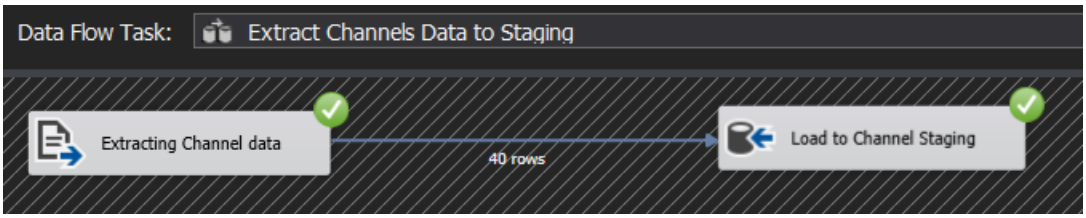
Extract Stores Data to Staging



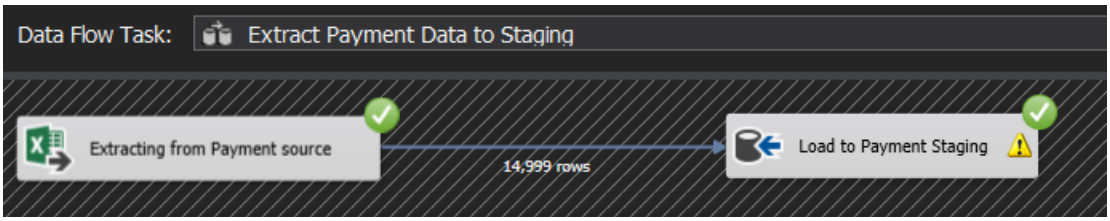
Extract Orders Data to Staging



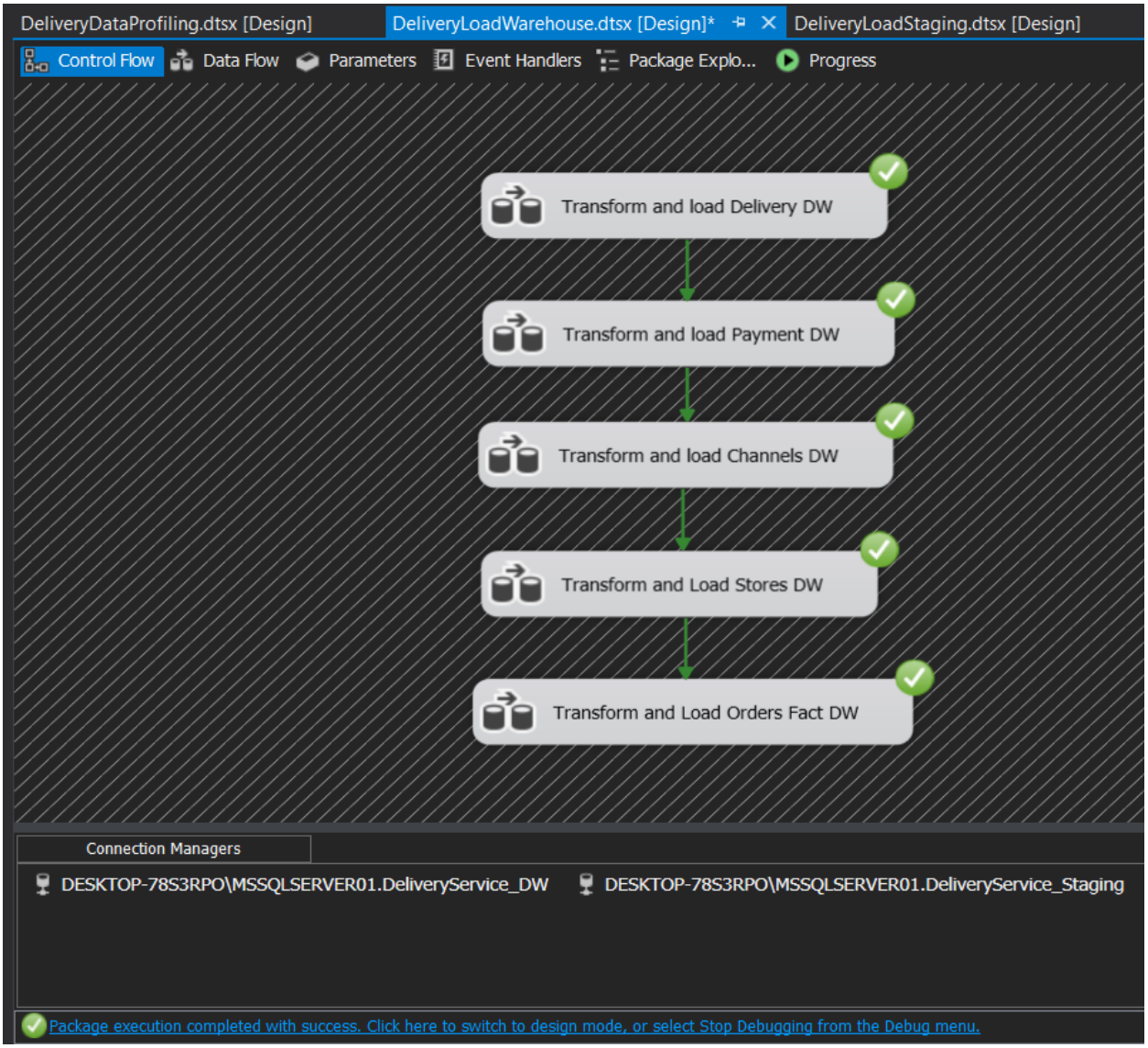
Extract Channels Data to Staging



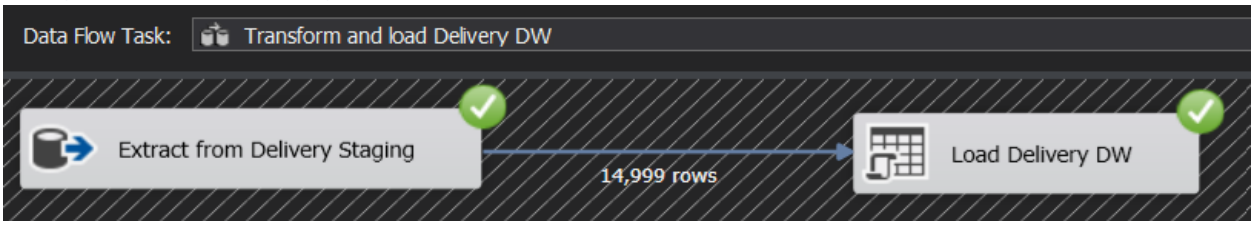
Extract Payments Data to Staging



Transforming and Loading to Data Warehouse



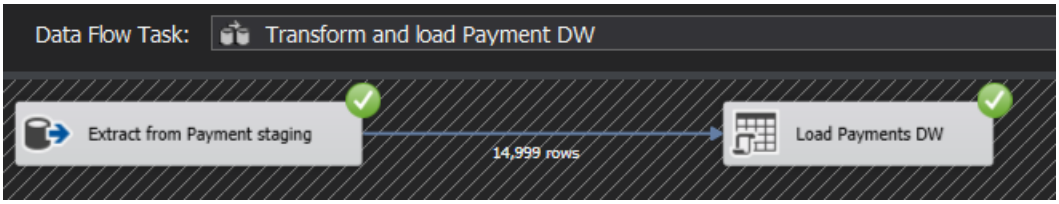
Transform and Load Delivery Dimension Table



UpdateDimDeliveries procedure was used here to update the existing values with the new values entered.

```
CREATE PROCEDURE dbo.UpdateDimDeliveries
@DeliveryID int,
@delivery_distance_meters int,
@delivery_status nvarchar(50)

AS
BEGIN
if not exists (select DeliverySK
from dbo.DimDeliveries
where AlternateDeliveryID = @DeliveryID)
BEGIN
insert into dbo.DimDeliveries
(AlternateDeliveryID, delivery_distance_meters, delivery_status,  InsertDate,
ModifiedDate)
values
(@DeliveryID, @delivery_distance_meters, @delivery_status,  GETDATE(), GETDATE())
END;
if exists (select DeliverySK
from dbo.DimDeliveries
where AlternateDeliveryID = @DeliveryID)
BEGIN
update dbo.DimDeliveries
set delivery_distance_meters = @delivery_distance_meters,
delivery_status = @delivery_status,
ModifiedDate = GETDATE()
where AlternateDeliveryID = @DeliveryID
END;
END;
```

Transform and Load Payments Dimension Table

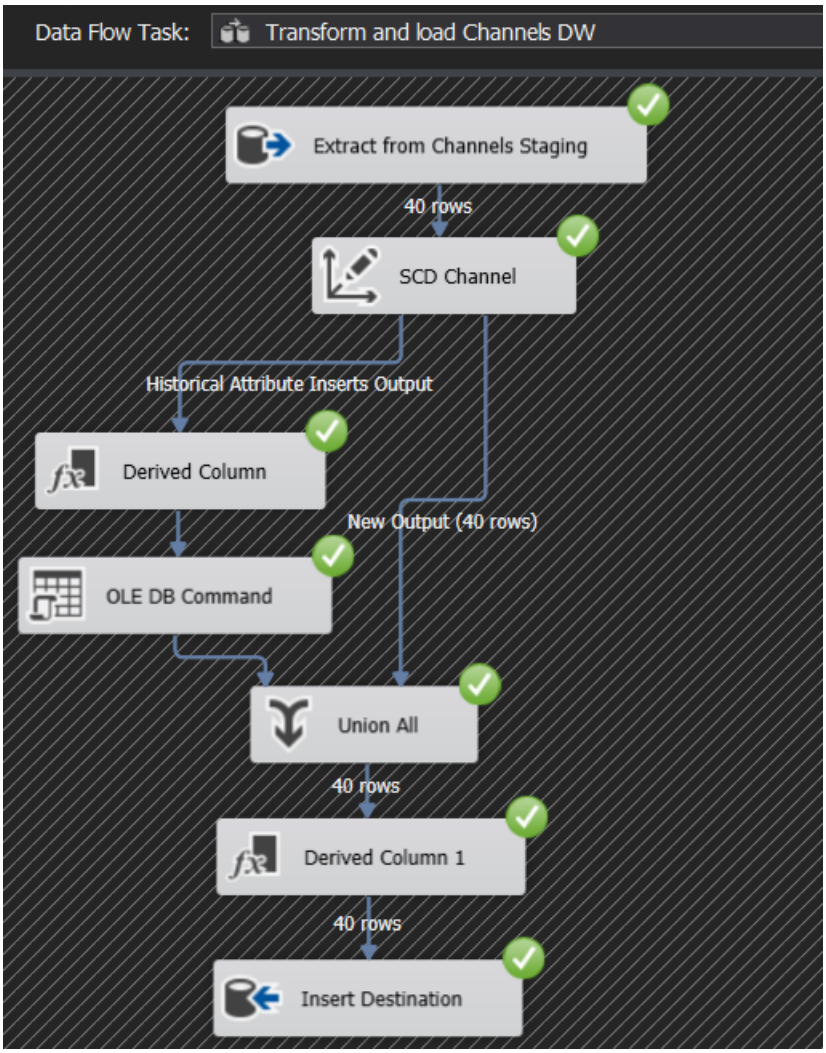


UpdateDimPayments procedure was used here to update the existing values with the new values entered.

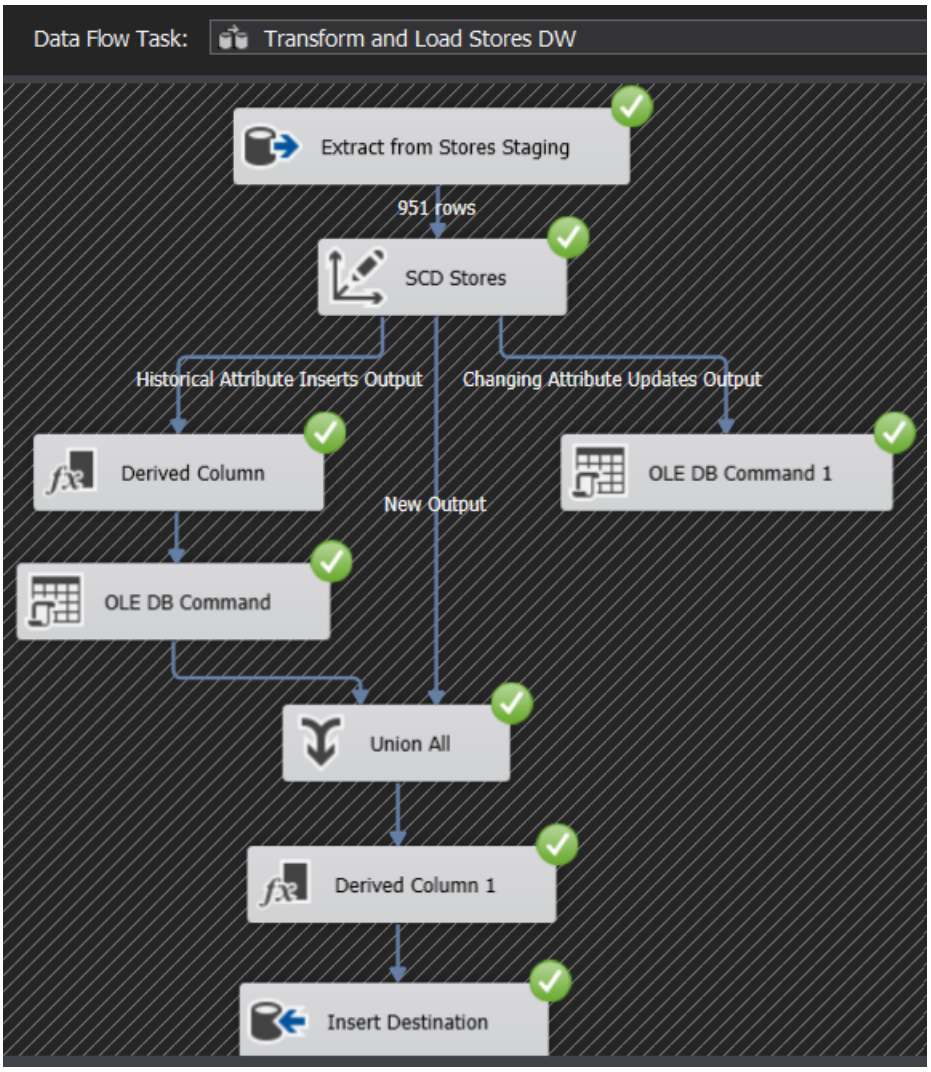
```
CREATE PROCEDURE dbo.UpdateDimPayments
@PaymentID int,
```

```
@payment_amount float,
@payment_fee float,
@payment_method nvarchar(50),
@payment_status nvarchar(50)

AS
BEGIN
if not exists (select PaymentSK
from dbo.DimPayments
where AlternatePaymentID = @PaymentID)
BEGIN
insert into dbo.DimPayments
(AlternatePaymentID, payment_amount, payment_fee, payment_method, payment_status,
InsertDate, ModifiedDate)
values
(@PaymentID, @payment_amount, @payment_fee, @payment_method, @payment_status,  GETDATE(),
GETDATE())
END;
if exists (select PaymentSK
from dbo.DimPayments
where AlternatePaymentID = @PaymentID)
BEGIN
update dbo.DimPayments
set payment_amount = @payment_amount,
payment_fee = @payment_fee,
payment_method = @payment_method,
payment_status = @payment_status,

ModifiedDate = GETDATE()
where AlternatePaymentID = @PaymentID
END;
END;
```

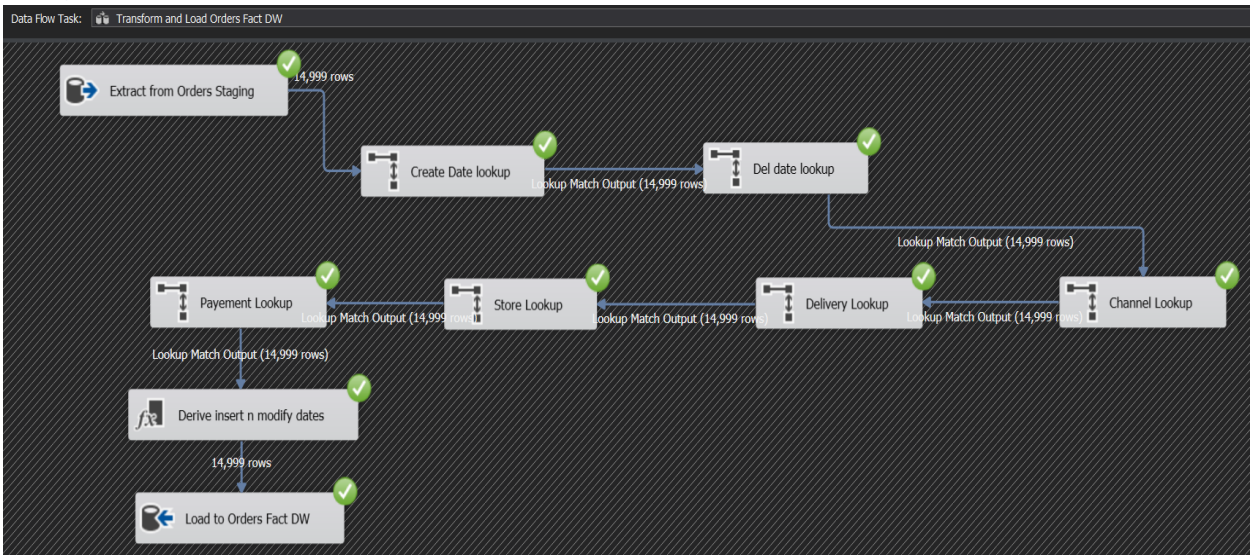
Transform and Load Channels Dimension Table



Transform and Load Stores Dimension Table



Transform and Load Orders Fact Table



Load DimDate Dimension Table

In the process I have used the DimDate Table that have been provided to get reference to 2 columns in the fact Table.

- created_date_key
- delivery_date_key

```
SELECT *
    FROM [DeliveryService_DW].[dbo].[DimDate]
```

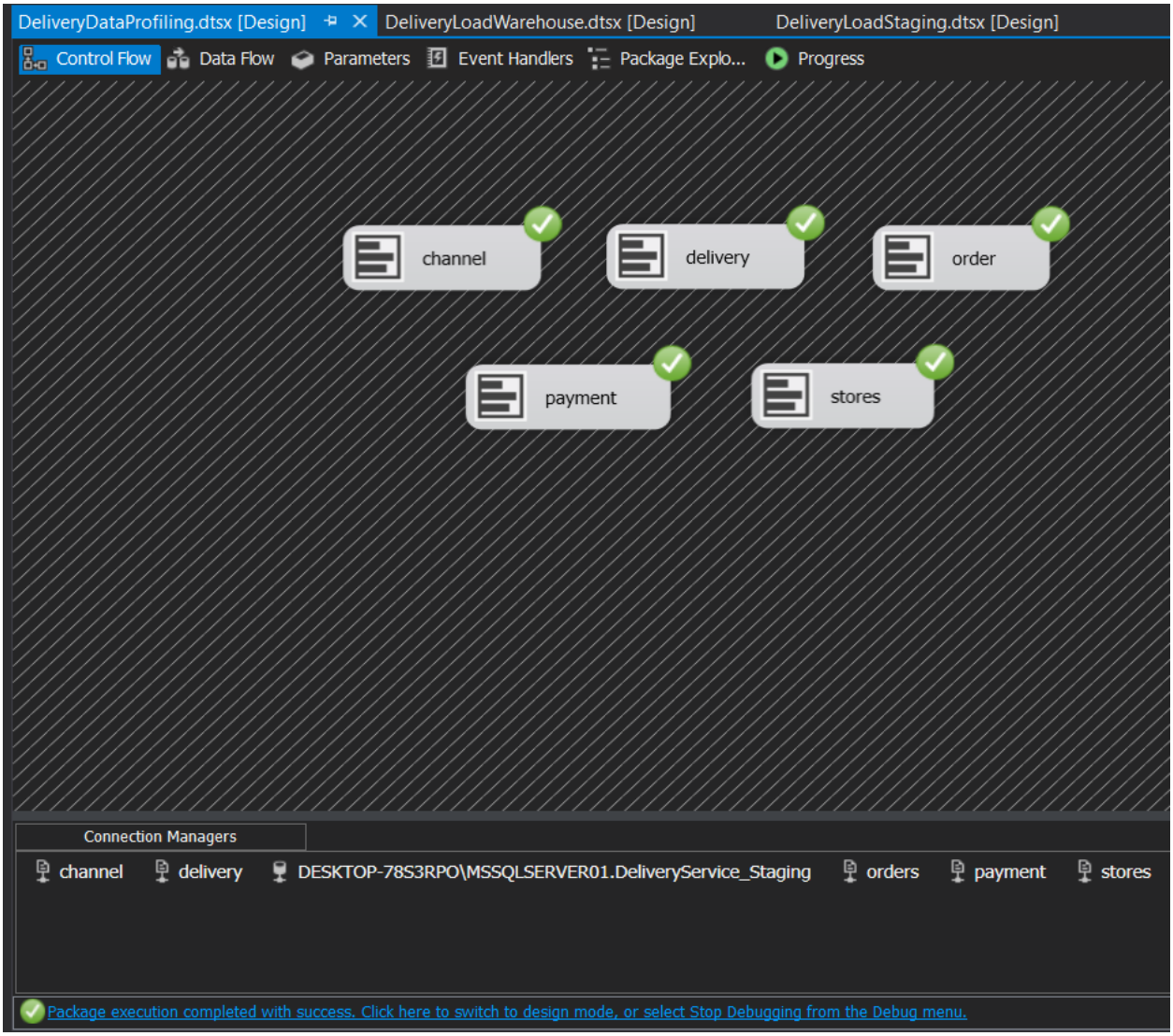
99 %

Results | Messages

	DateKey	Date	FullDateUK	FullDateUSA	DayOfMonth	DaySuffix	DayName	DayOfWeekUSA	DayOfWeekUK	DayOfWeekInMonth	DayOfWeekInYear	DayOfQuarter	DayOfYear	WeekOfMonth	WeekOfQuarter	Weel
1	19900101	1990-01-01 00:00:00.000	01/01/1990	01/01/1990	1	1st	Monday	2	1	1	1	1	1	1	1	1
2	19900102	1990-01-02 00:00:00.000	02/01/1990	01/02/1990	2	2nd	Tuesday	3	2	1	1	1	2	1	1	1
3	19900103	1990-01-03 00:00:00.000	03/01/1990	01/03/1990	3	3rd	Wednesday	4	3	1	1	1	3	1	1	1
4	19900104	1990-01-04 00:00:00.000	04/01/1990	01/04/1990	4	4th	Thursday	5	4	1	1	1	4	1	1	1
5	19900105	1990-01-05 00:00:00.000	05/01/1990	01/05/1990	5	5th	Friday	6	5	1	1	1	5	1	1	1
6	19900106	1990-01-06 00:00:00.000	06/01/1990	01/06/1990	6	6th	Saturday	7	6	1	1	1	6	1	1	1
7	19900107	1990-01-07 00:00:00.000	07/01/1990	01/07/1990	7	7th	Sunday	1	7	1	1	1	7	2	1	2
8	19900108	1990-01-08 00:00:00.000	08/01/1990	01/08/1990	8	8th	Monday	2	1	2	2	2	8	2	2	2
9	19900109	1990-01-09 00:00:00.000	09/01/1990	01/09/1990	9	9th	Tuesday	3	2	2	2	2	9	2	2	2
10	19900110	1990-01-10 00:00:00.000	10/01/1990	01/10/1990	10	10th	Wednesday	4	3	2	2	2	10	2	2	2

Data Profiling

ETL data profiling is a detailed analysis of source data. It attempts to identify the structure, quality, and content of the source data, as well as its relationships with other data. It has been used during the extraction, transformation, and loading (ETL) process.



References

https://docs.microsoft.com/en-us/sql/sql-server/end-of-support/sql-server-end-of-support-overview?view=sql-server-ver16

https://www.w3schools.com/sql/

https://stackoverflow.com/

https://www.toolbox.com/tech/data-management/question/is-it-necessary-to-truncate-table-before-running-loading-mapping-workflow-021012/

Thank You