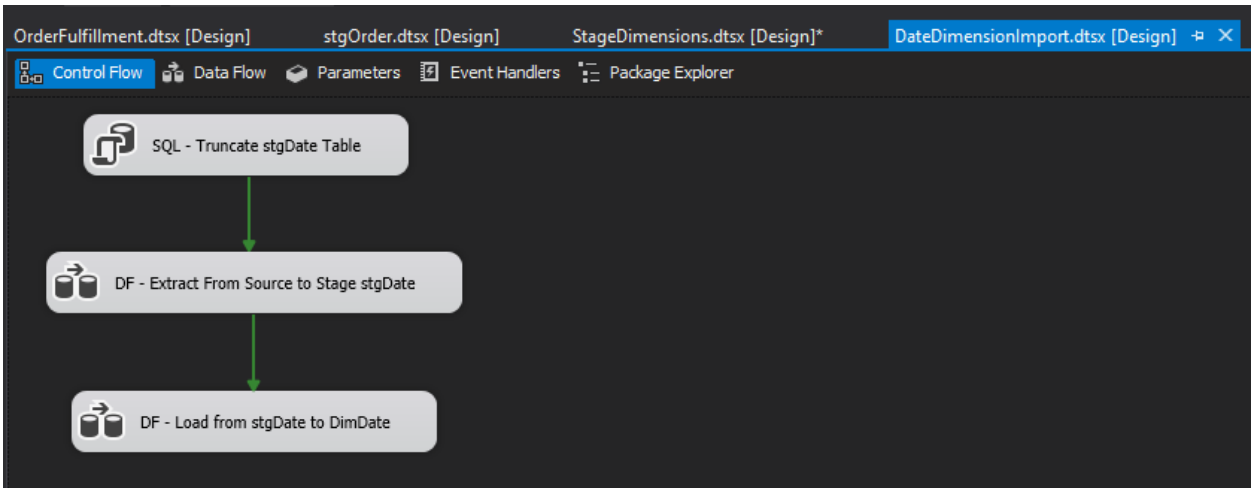
Project Document: ETL Process
IST 722 – Data Warehouse | 01/14/2022

Daniel Caley | Michael Johnson | Samuel Deery-Schmitt | Jennifer Lammers Zimmer

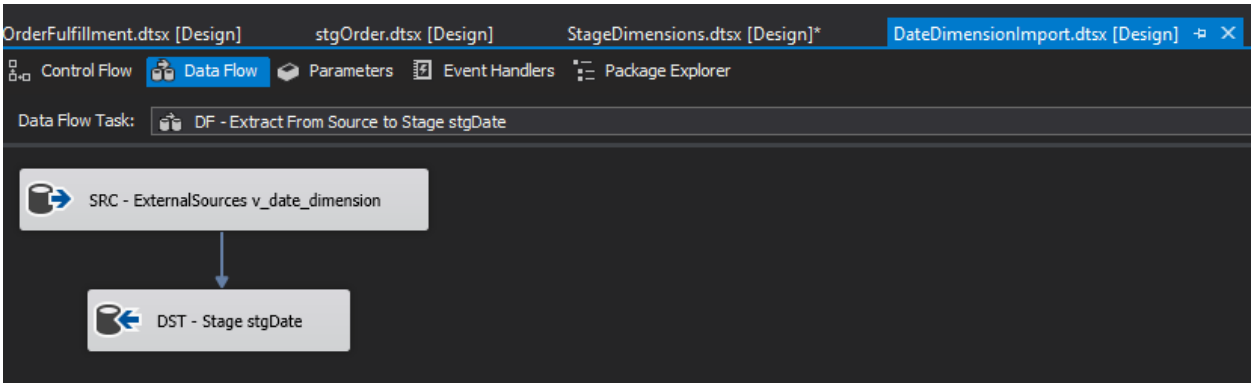
Contents

Date Dimension: Full ETL

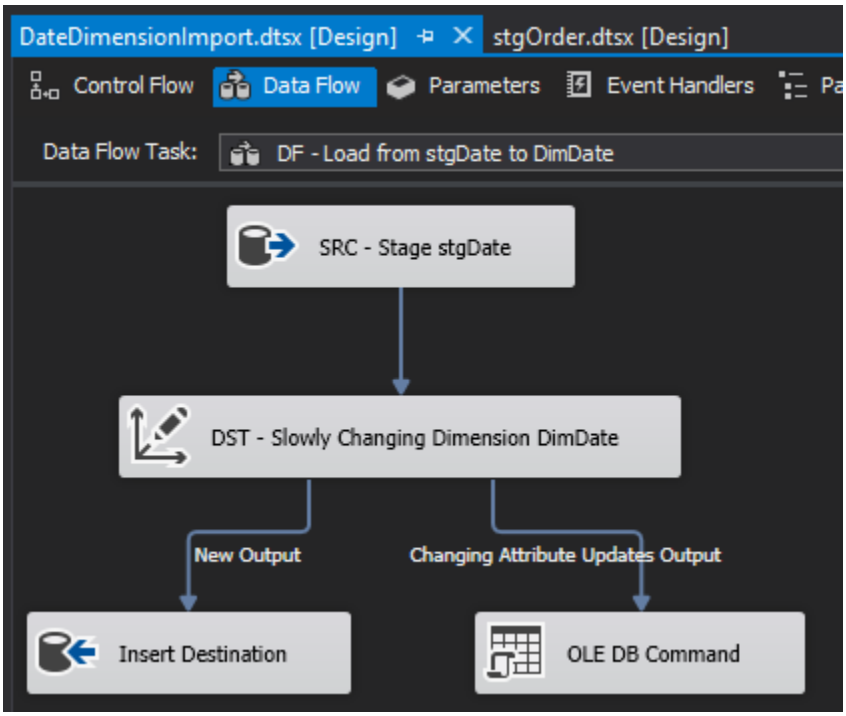
Control Flow: For this dimension, we performed ETL as a single control flow; for other dimensions, we loaded into the data warehouse in a separate control flow.



Data Flow: This is the first data flow in the control flow, in which we pull date fields from the ExternalSources2 schema and place into the staging schema.

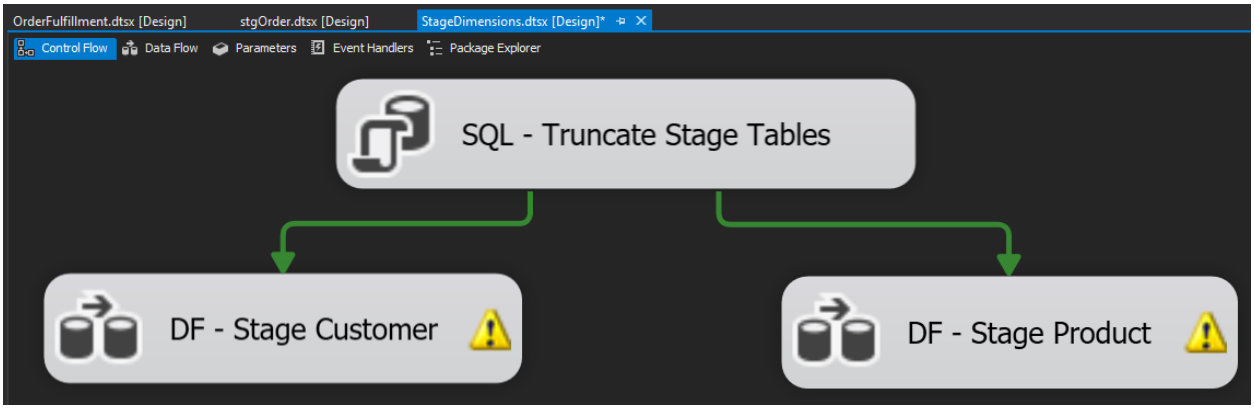


Data Flow: In the second data flow for the DimDate, we load the date data from the staging schema to the data warehouse, treating it as a Type 1 SCD.

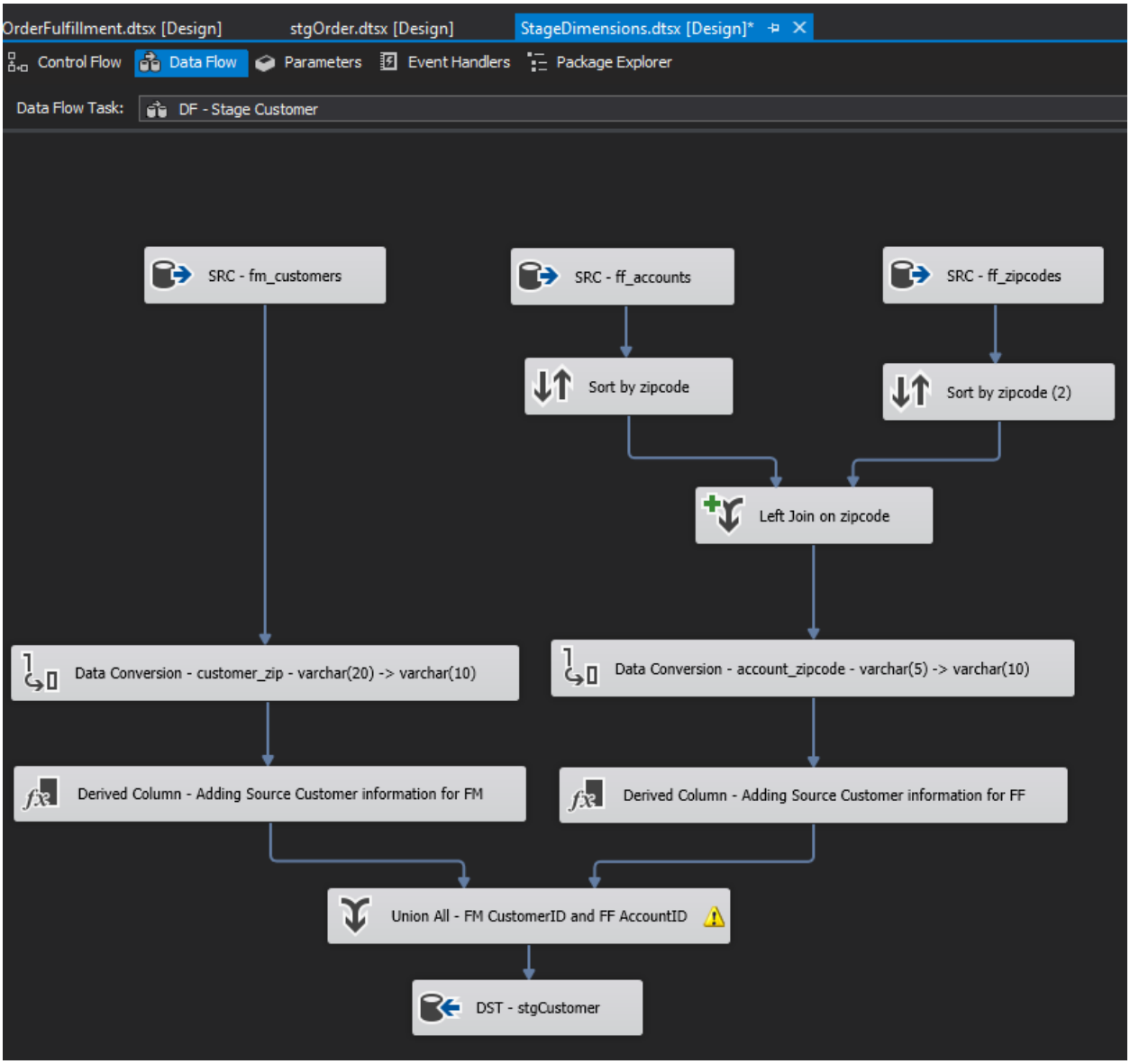


Staging Customer and Product Dimensions

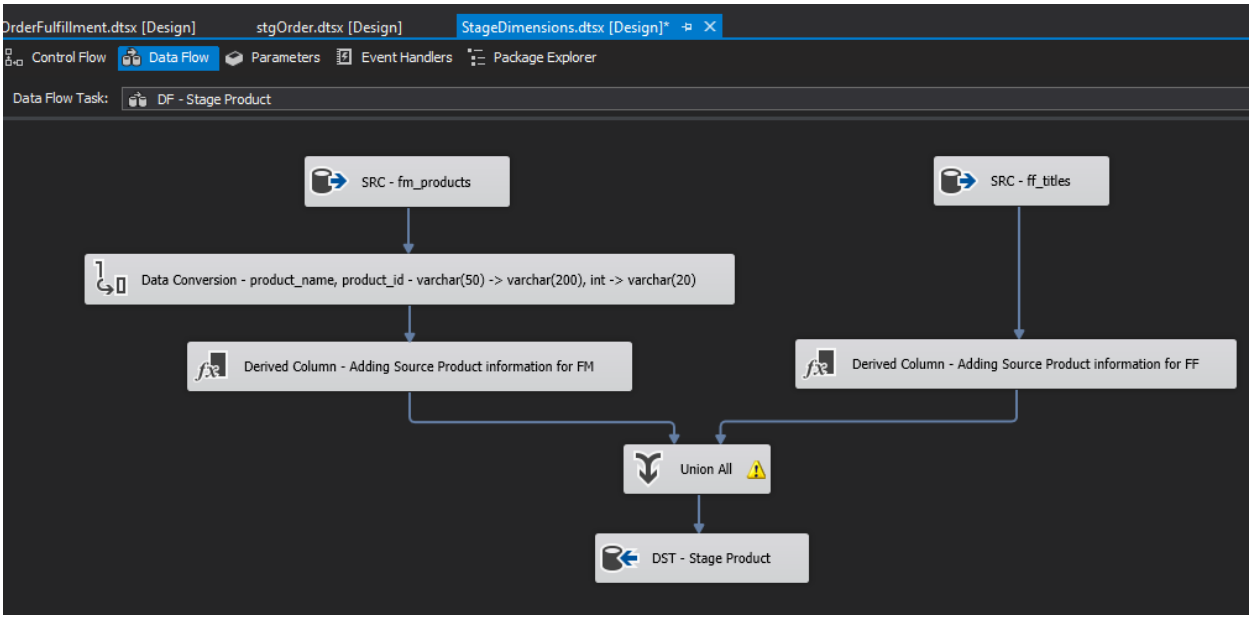
Control Flow: This control flow generates the Staging tables for the Customer and the Product Dimensions.



Customer Data Flow: Merged Fudgemart and Fudgeflix customer data into one customer table. Resolved data type mismatch for zip codes between databases with data conversion. Created a column in the merged data table to show original source of customer data (Fudgemart or Fudgeflix).

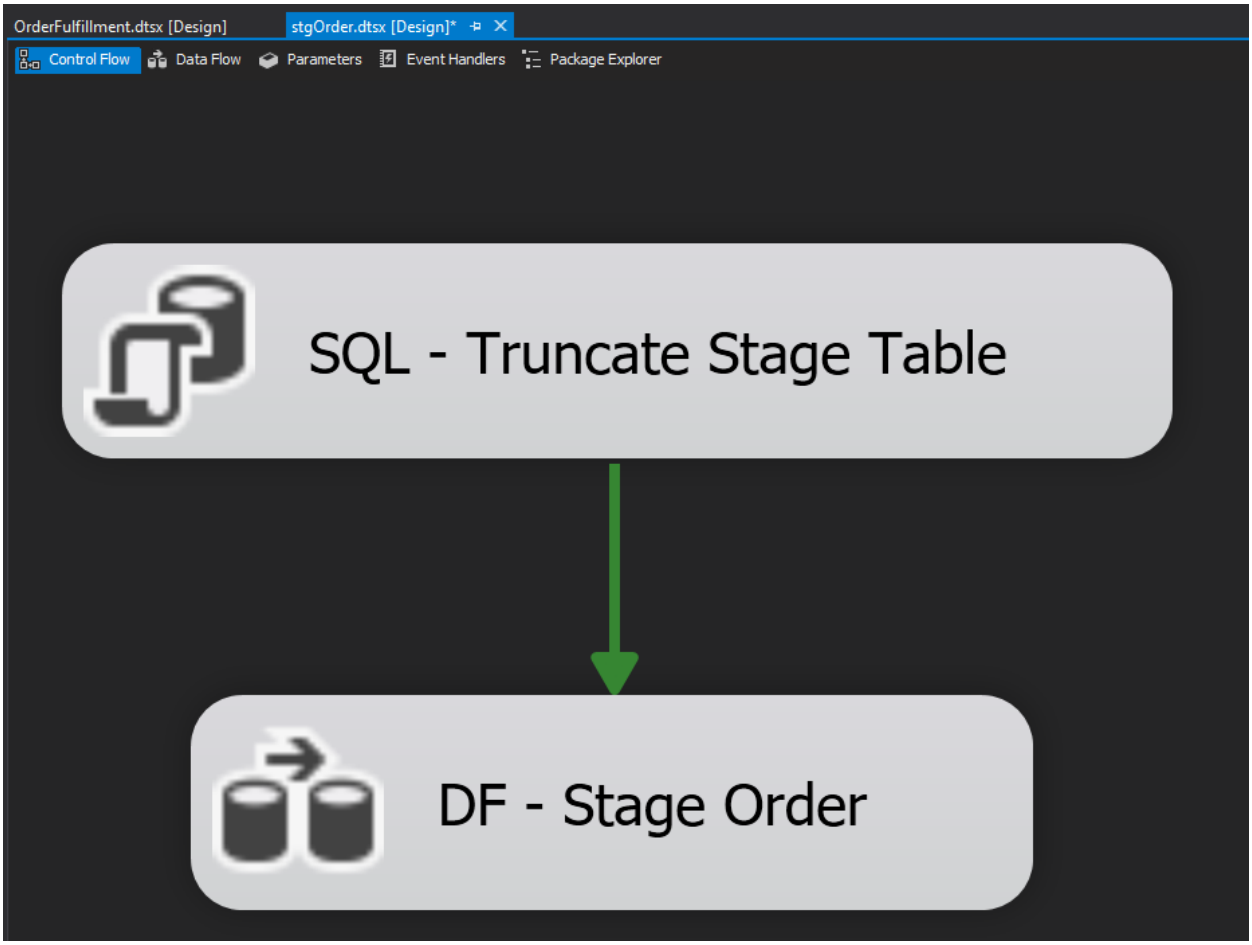


Product Data Flow: Merged Fudgemart products and Fudgeflix titles into one data table. Data conversion on Fudgemart product_name, product_id to match data types in Fudgeflix. Created a derived column for both data sources to indicate orginal data source in final Staging table.



Staging Order

Control Flow: This control flow builds a single table, Order, in Stage.



Order Data Flow: Building the stage order required bringing in sources from both Fudgemart and Fudgeflix related to their respective orders and order details. For Fudgemart, the schema required just order and order details and the schema for fudgeflix required just the account titles table. Since the Customer and Product stage tables were already created with the original keys and a column for the original databases, the data flow also included a source connection to the two stage tables with queries to pull the Fudgemart or Fudgeflix data separately.

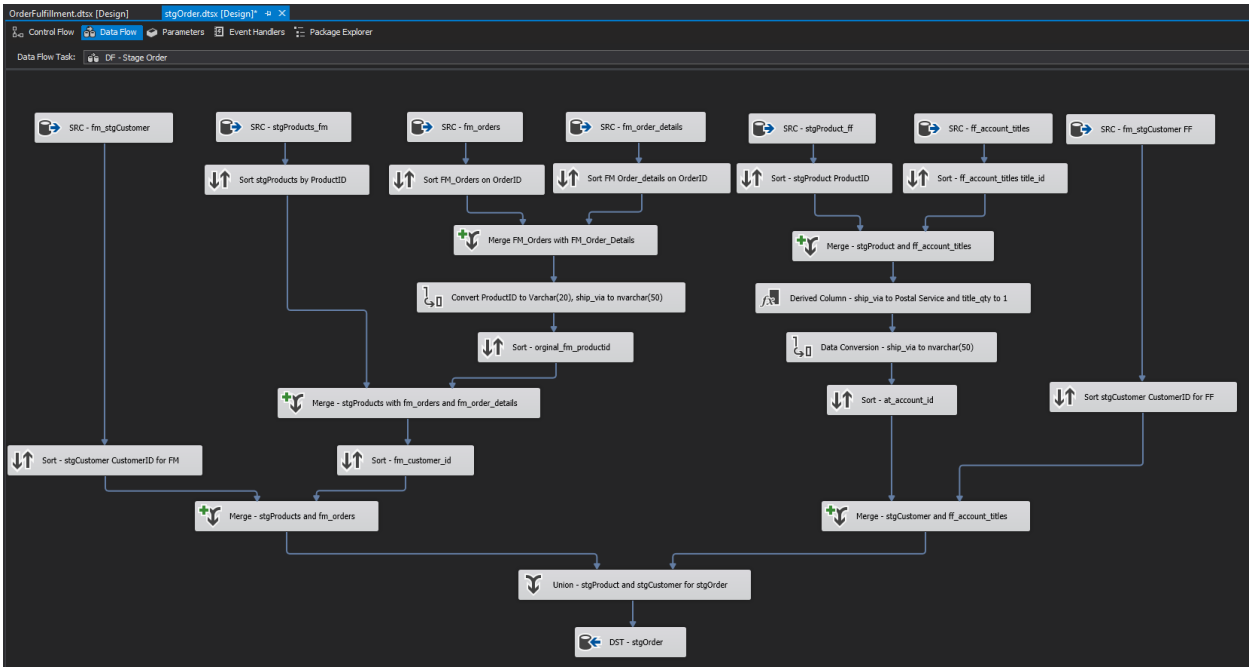
Fudgemart side:

Fudgemart orders and order details were merged together based on order ID, with product ID converting to varchar(20) to match with the product ID from Fudgeflix. The ship via column was also changed to nvarchar(50). The next merge on the Fudgmart side was between the orders and staged product table, with a final merge on the staged customer table. This ensures the surrogate primary keys are mapped to the correct order and customer.

Fudgeflix side:

Fudgeflix account titles table contains the order details and was merged with the stage product table to correctly map the surrogate key with the original account title ID. Ship via was added to the merged table with “Postal Service” as the default value for all Fudgeflix orders and a quantity column was also added with a default “1” for every order. Ship via was converted to nvarchar(50) to match with Fudgemart and then was merged with the staged customer table to map the correct surrogate customer key with the original customer who placed the order.

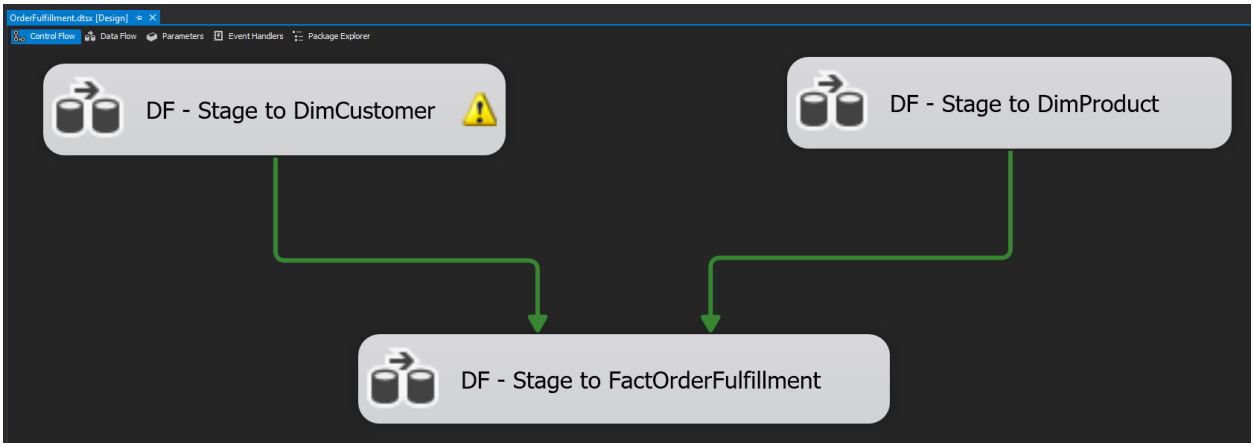
The final union joined the Fudgemart and Fudgeflix databases with the correct data types and associated attributes required for the final fact order fulfillment table.



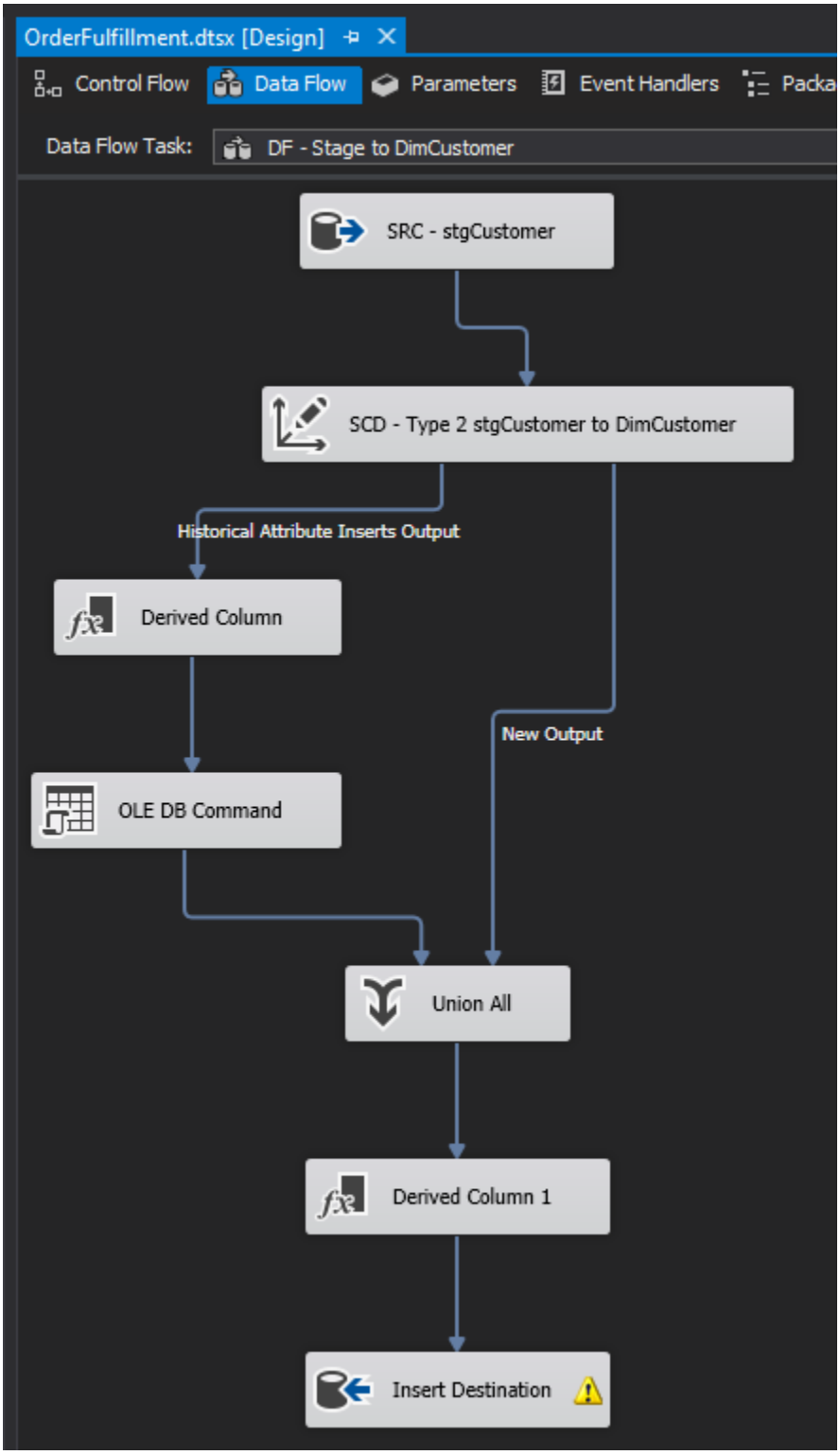
Load Dimension & Fact Table

Load Customer Dimension, Product Dimension, and Fact Table from Stage to Data Warehouse

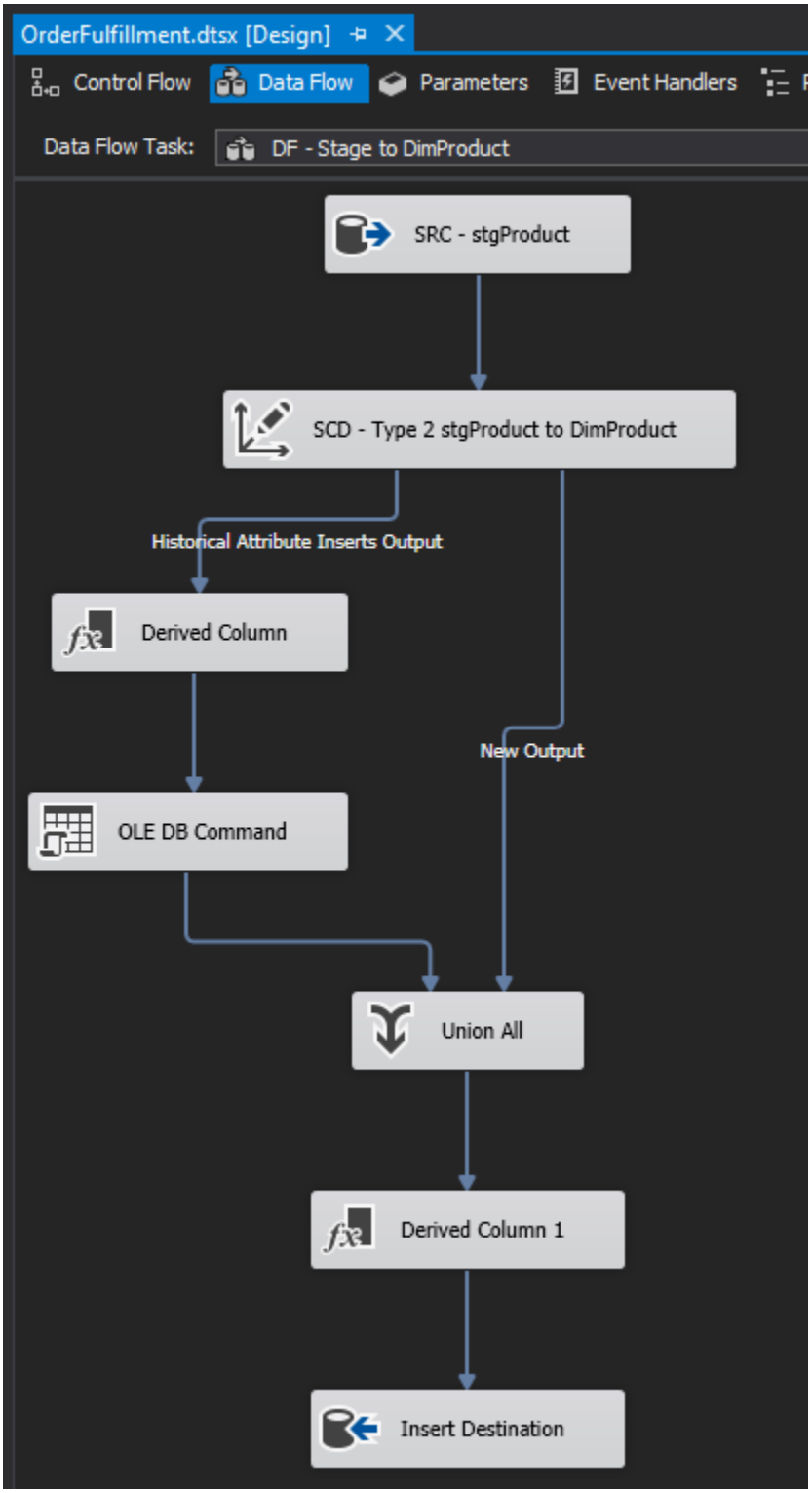
Control Flow: This control flow moves the transformed data from Stage to the Data Warehouse for the Customer Dimension, the Product Dimension and the Fact Table.



Slowly Changing Dimension Customer Data Flow: To create the Customer Dimension, slowly changing dimension attributes are defined for each field in the stgCustomer table. All attributes in the Customer Dimension are set to Type 2 upsert.



Slowly Changing Dimension Product Data Flow: To create the Product Dimension, slowly changing dimension attributes are defined for each field in the stgProduct table. All attributes in the Product Dimension are set to Type 2 upsert.



Fact OrderFulfillment Data Flow: The data flow for fact order fulfillment started with the staged order data that did most of the heavy lifting to tie customers, products, and orders together. Two lookups connected the Product Key and the Customer Key to the Fact Table and another two lookups found the Order Date Key and the Shipped Date Key. One derived column containing the measure “Order to ship lag in days” was added to the fact table. The final table was then sent to the data warehouse.