Sri Lanka Institute of Information Technology

Data Warehouse and Business Intelligence

IT3021

3rd Year, 1st Semester

Assignment 1

Weekday Batch
Y3S1.15(DS)

IT19021430
Hillary J.R.

Contents

Contents

Data Set Selection

The selected data source is a collection of transactional data. The link to the source data set is mentioned below:

Link to chosen data set
https://www.kaggle.com/ghoshsaptarshi/av-genpact-hack-dec2018

The select data set is based on a meal delivery company which operates in multiple cities. The data set consists of various fulfillment centers in these cities for dispatching meal orders to their customers. Through the data set can be used to help these centers with demand forecasting for upcoming weeks so that these centers will plan the stock of raw materials accordingly.

Aim of the data set
The source data set is been provided to predict the demand for the next 10 weeks based on the history of 145 weeks for the center meal combinations.
Staffing of centers based on demand
Procurement planning – Raw materials (raw materials are perishable)

The source data set consists

- Historical data of demand for a product-center combination (Weeks: 1 to 145)
- Product (Meal) features such as category, sub-category, current price and discount
- Information for fulfillment center like center area, city information etc.

1. **train.csv**

Weekly Demand data: Contains the historical demand data for all centers.

Variable	**Definition**
id	Unique ID
week	Week number
center_id	Unique ID for fulfillment center
meal_id	Unique ID for meal
checkout__price	Final price including discount, taxes & delivery charges
base_price	Base price of the meal – this includes profit margin
emailer_for_promotion	E-Mailer sent for promotion
homepage_featured	Meal featured at homepage
num_orders	Number of orders sold per meal per center

2. **fulfilment_center_info.csv**

Contains information for each fulfilment center

Variable	**Definition**
center_id	Unique ID for fulfillment center
city_code	Unique code for city
region_code	Unique code for region
center_type	Anonymized center type
op_area	Area of operation (in km^2)

3. **meal_info.csv**

Contains information for each meal being served

Variable	Definition
meal_id	Unique ID for the meal
category	Type of meal (beverages/snacks/soups...)
cuisine	Meal cuisine (Indian/Italian/...)

Preparation Of Data Sources

Modifications were done accordingly to the data set derived from the source. Although the source contains only .csv files I have made some changes to the source files and to match the assignment specifications. According to the changes I made my data source contains of three types such .csv files, .txt files and .bak

Assumptions
Week number was changed into a date considering the week 1 as first week of the year 2018 and the 7th day of that week was considered to be the day data was loaded to the source tables. This assumption was taken to reduce the complexity and to make it easier to look up the DimDate table when loading to the data ware house. Each unique Meal ID was additionally given a Meal Name which was part from the source data to understand the variations easily when analyzing rather than analyzing using numeric meal ID values.
SrcCenterDetails, SrcCenterManager, SrcCenterManagerDetails were additionally taken(derived) data apart from the source to match the assignment specifications and increase the complexity of the scenario.

Source table details

- SrcMeal.CSV
- SrcMealBeverage.CSV
- SrcMealCuisine.CSV
- MealDemand_SourceDB.bak – SrcCenter, SrcCenterDetails
- SrcCity.txt
- SrcRegion.txt
- SrcCenterManager.txt
- SrcCenterManagerDetails.csv

Further details about the tables, attributes and datatypes of each attribute are given in the table below.

Source: SrcMeal.CSV		Source Type: CSV File		Table Name: SrcMeal
Column Name	**Data Type**	**Link Table**	**Link Column**	**Description**
MealID	Int			Unique ID.
MealName	Nvarchar(50)			Meal Name
MealCuisineID	Int	SrcMealCuisine	MealCuisineID	Meal Cuisine
MealBeverageID	Int	SrcMealBeverage	MealBeverageID	Meal Beverage
ModifiedDate	DateTime			Modified Date of the Meal
Source: SrcMealBeverage.CSV		Source Type: CSV File		Table Name: SrcMealBeverage
Column Name	**Data Type**	**Link Table**	**Link Column**	**Description**

MealBeverageID	Int			Unique ID.
MealBeverageName	Nvarchar(50)			Name of the Meal Beverage
ModifiedDate	DateTime			Modified Date of the Meal Beverage

Source: SrcMealCuisine.CSV	Source Type: CSV File		Table Name: SrcMealCuisine

Column Name	**Data Type**	**Link Table**	**Link Column**	**Description**
MealCuisineID	Int			Unique ID.
MealCuisineName	Nvarchar(50)			Name of the Meal Cuisine
ModifiedDate	DateTime			ModifiedDate of the Meal Cuisine

Source: MealDemand_SourceDB	Source Type: SQL Database		Table Name: SrcCenter

Column Name	**Data Type**	**Link Table**	**Link Column**	**Description**
CenterID	Int			Center Unique ID.
CityID	Int	SrcCity	CityID	ID of the city
CenterManagerID	Int	SrcCenterManager	CenterManagerID	ID of the Center Manager
CenterType	Nvarchar(50)			Anonymized center type
OpperationaArea	Nvarchar(50)			Area of operation(km^2)

Source: MealDemand_SourceDB	Source Type: SQL Database		Table Name: SrcCenterDetails

Column Name	**Data Type**	**Link Table**	**Link Column**	**Description**
CenterID	Int			Center Unique ID.
PhoneNumber	Nvarchar(25)			Phone number of the Center
PhoneNumberType	Nvarchar(50)			Phone type of the Center
NumberOfEmployees	Int			Number of employees working in a particular center

Source: SrcCity.txt	Source Type: Text File		Table Name: SrcCity

Column Name	**Data Type**	**Link Table**	**Link Column**	**Description**
CityID	Int			Unique ID.
CityCode	Int			Unique Code of the region
RegionID	Int	SrcRegion	RegionID	ID of the region
ModifiedDate	DateTime			ModifiedDate of the City

Source: SrcRegion.txt	Source Type: Text File		Table Name: SrcRegion

Column Name	**Data Type**	**Link Table**	**Link Column**	**Description**
RegionID	Int			Unique ID.
RegionCode	Int			Unique Code of the region
ModifiedDate	DateTime			ModifiedDate of the Region

Source: SrcCenterManager.txt	Source Type: Text File		Table Name: SrcCenterManager

Column Name	**Data Type**	**Link Table**	**Link Column**	**Description**
CenterManagerID	Int			Unique ID.
Title	Nvarchar(8)			Center Manager Title (Mr., Mrs etc.)
FirstName	Nvarchar(50)			First name of the Center Manager
LastName	Nvarchar(50)			Last name of the Center Manager
Gender	Nvarchar(1)			Center Manager Gender (M, F)

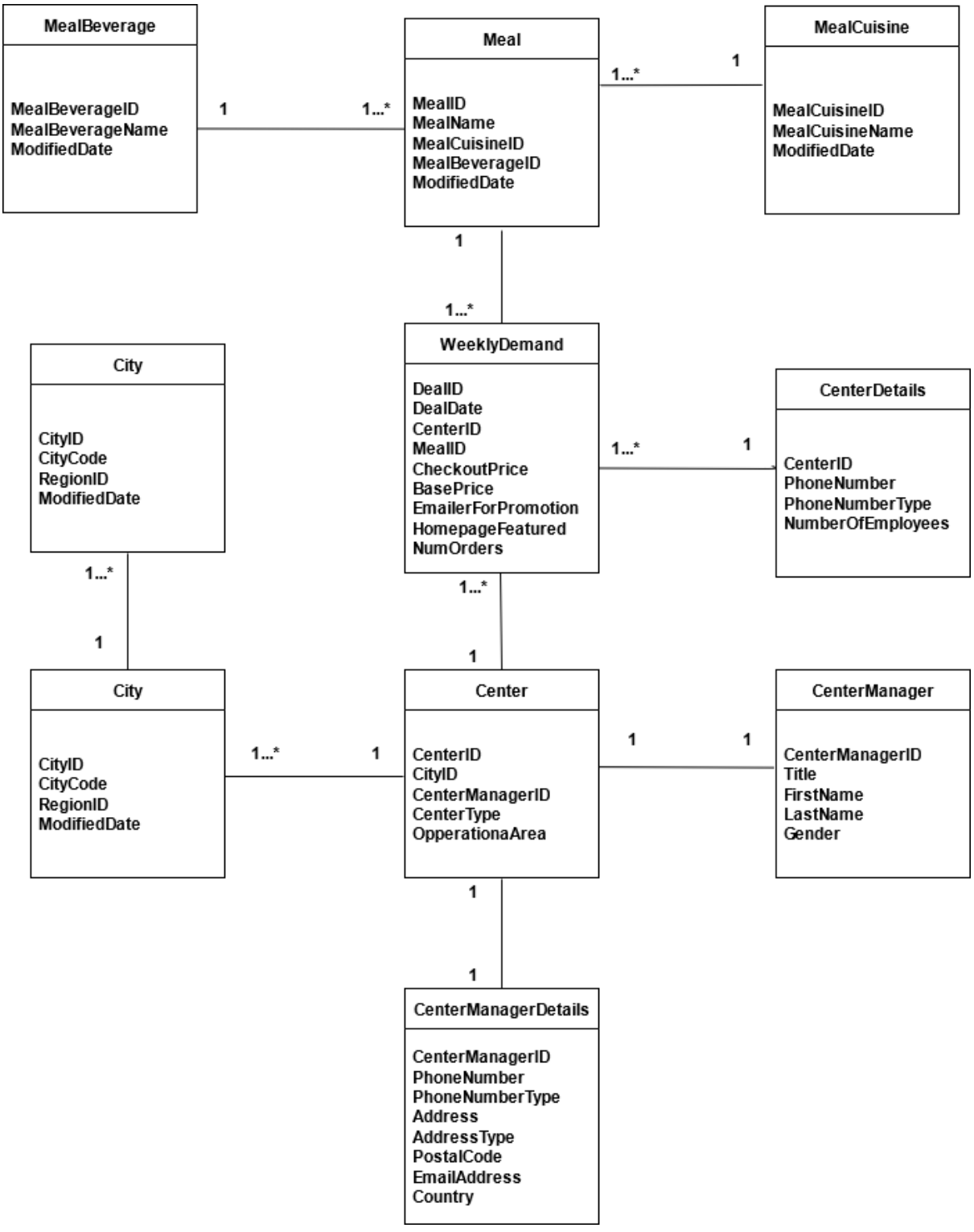
Source: SrcCenterManagerDetails.csv	Source Type: CSV File		Table Name: SrcCenterManagerDetails

Column Name	Data Type	Link Table	Link Column	Description
CenterManagerID	Int			Unique ID.
PhoneNumber	Nvarchar(25)			Phone number of the Center Manager
PhoneNumberType	Nvarchar(50)			Phone type of the Center Manager
Address	Nvarchar(50)			Address of the Center Manager
AddressType	Nvarchar(50)			Address type of the Center Manager
PostalCode	Nvarchar(50)			Postal code of the Center Manager
EmailAddress	Nvarchar(50)			Email address of the Center Manager
Country	Nvarchar(50)			Country name of the Center Manager

Source: MealDemand_SourceDB	Source Type: SQL Database	Table Name: SrcWeeklyDemand

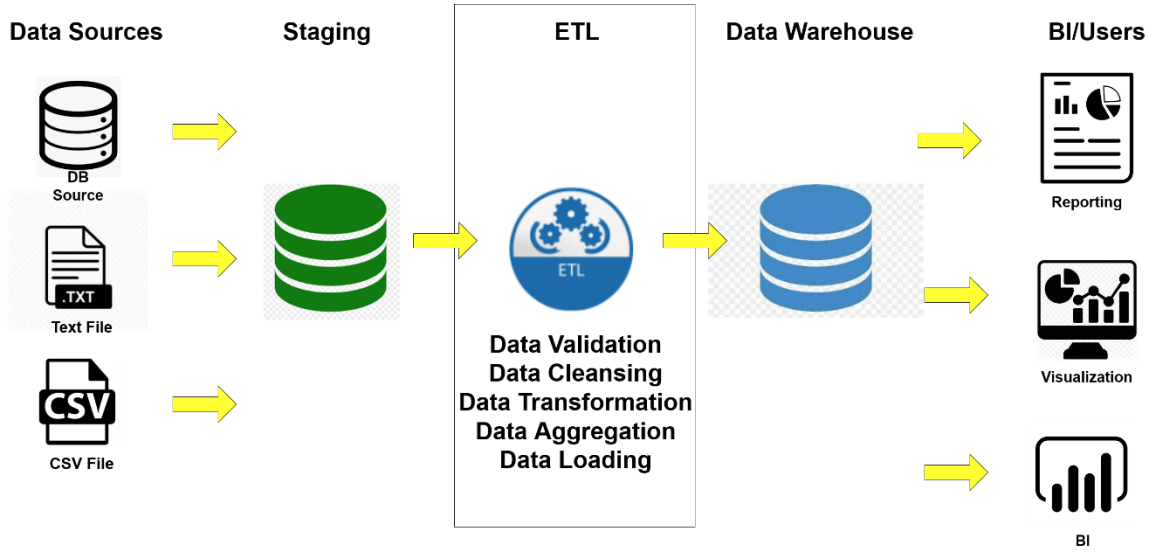
Column Name	Data Type	Link Table	Link Column	Description
DealID	Int			Unique ID.
DealDate	DateTime			7th Day date of week
CenterID	Int	SrcCenter	CenterID	ID of the Center Manager
MealID	int	SrcMeal	MealID	ID of the Meal
CheckoutPrice	Money			Final price inlcuding discount, taxes & delivery charges
BasePrice	Money			Base price of the meal
EmailerForPromotion	Int			Emailer sent for promotion of the meal
HomepageFeatured	Int			Meal featured at home page
NumOrders	Int			Target(Orders Count)

Class Diagram using Source tables



High-Level BI Solution Architecture

The basic concept of a Data Warehouse is to facilitate a single version of truth for a company for decision making and forecasting. A Data warehouse is an information system that contains historical and commutative data from single or multiple sources. Data Warehouse Concepts simplify the reporting and analysis process of organizations.



Data Sources
This represents the different data sources that feed data into the data warehouse. The data source can be of any format plain text file, relational database, other types of database, Excel file, etc., can all act as a data source. Data sources are the locations where data is being used come from.

For the given scenario, primary data source is a database and secondary data sources are, a flat file and a csv file.

Data Extraction Layer
Data gets pulled from the data source into the data staging layer. There is likely some minimal data cleansing, but there is unlikely any major data transformation.

Staging Area
This is where data will be gathered prior to being taken and transformed into a data warehouse. Having one common area makes it easier for subsequent data processing further for the data warehouse.

ETL
ETL stands for Extract, Transform and Load. This is where data gains its importance, as logic is applied to transform the data from a transactional nature to an analytical nature. It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area and then finally, loads it into the Data Warehouse system.

1. Extraction – reading of source data/ In this case staging layer data which is something similar to the source data but only difference is everything is taken into a common format and common place
2. Transformation -preparing data to be inserted to the target model, this includes cleansing, integrating, de-duplication, enriching, aggregation and loading.
3. Loading - The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse. Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals.

Data Warehouse

This is where the transformed and cleansed data sit. Based on scope and functionality, 3 types of entities can be found here: data warehouse, data mart, and operational data store (ODS). In any given system, you may have just one of the three, two of the three, or all three types. In our scenarios we have loaded the data into facts and dimensional tables. DWs are central repositories of integrated data from one or more sources.

BI Layer

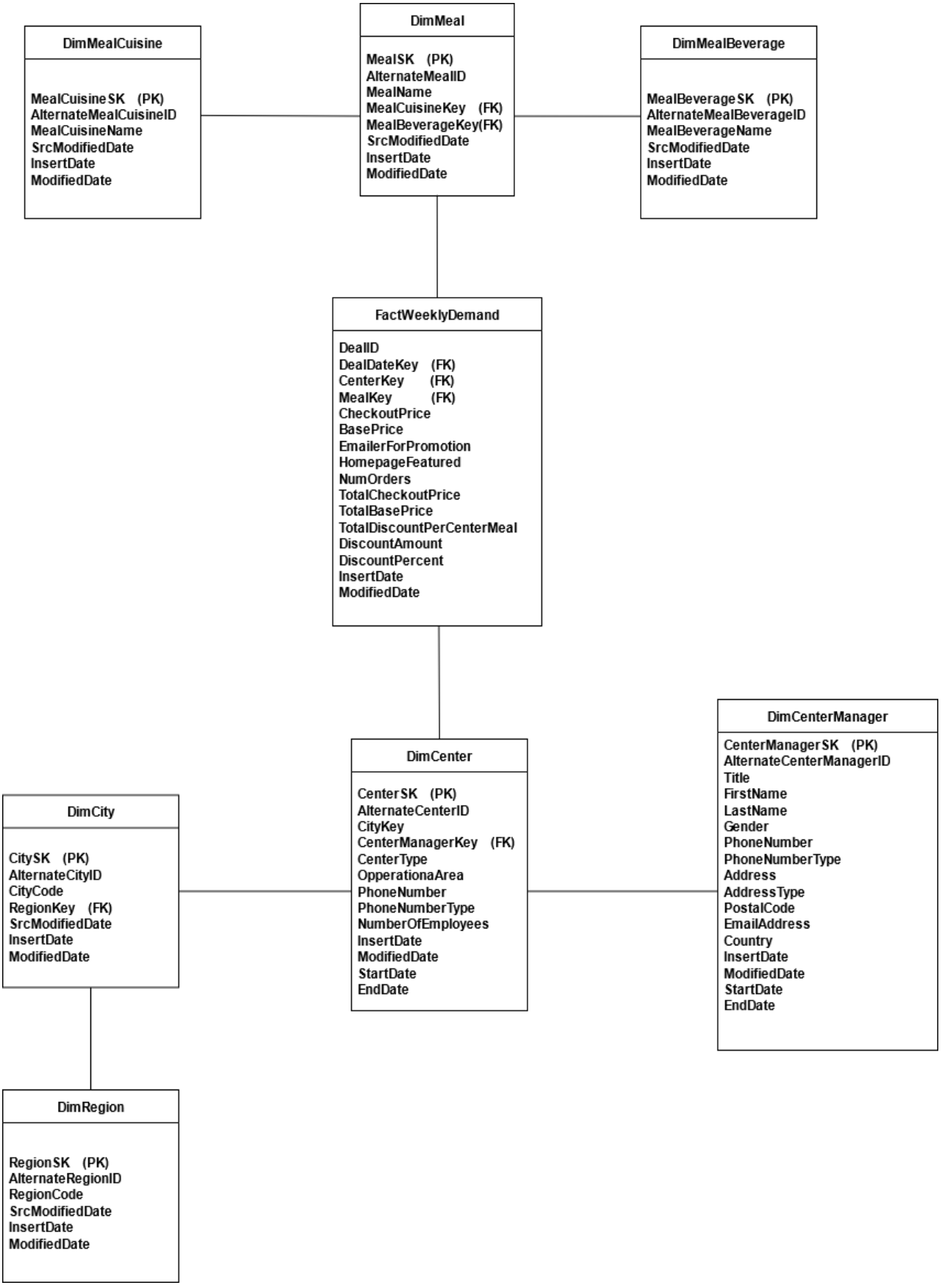
In business intelligence, data warehouses serve as the backbone of data storage. Business intelligence relies on complex queries and comparing multiple sets of data to inform everything from everyday decisions to organization-wide shifts in focus.

This layer includes

BI Applications - web applications, mobile applications, self-service BI tools, other data mining and modelling tools.

Data warehouse design & development

Snow Flake Schema was the chosen schema for the data ware house development based on the data set
In this schema, it shows the final structure of the data warehousing for the scenario Meal Demand Forecast data set. Here, redundancy will not occur, so the efficiency of storage is improved.



Dimension Name	Dimension Attributes	Derived	DataType		Key	Derived Logic
DimMeal	**MealSK**	Y		Not Null	PK	Auto incrementing
	AlternateMealID	N	Int	Not Null		
	MealName	N	Nvarchar(50)			
	MealCuisineKey	N	Int		FK	
	MealBeverageKey	N	Int		FK	
	SrcModifiedDate	N	DateTime			
	InsertDate	Y	datetime			SysDateTime
	ModifiedDate	Y	datetime			SysDateTime
DimMealCuisine	**MealCuisineSK**	Y	Int	Not Null	PK	Auto incrementing
	AlternateMealCuisineID	N	Int	Not Null		
	MealCuisineName	N	Nvarchar(50)			
	SrcModifiedDate	N	DateTime			
	InsertDate	Y	datetime			SysDateTime
	ModifiedDate	Y	datetime			SysDateTime
DimMealBeverage	**MealBeverageSK**	Y	Int	Not Null	PK	Auto incrementing
	AlternateMealBeverageID	N	Int	Not Null		
	MealBeverageName	N	Nvarchar(50)			
	SrcModifiedDate	N	DateTime			
	InsertDate	Y	datetime			SysDateTime
	ModifiedDate	Y	datetime			SysDateTime
DimCenterManager	**CenterManagerSK**	Y	Int	Not Null	PK	Auto incrementing
	AlternateCenterManagerID	N	Int	Not Null		
	Title	N	Nvarchar(8)			
	FirstName	N	Nvarchar(50)			
	LastName	N	Nvarchar(50)			
	Gender	N	Nvarchar(1)			
	PhoneNumber	N	Nvarchar(25)			
	PhoneNumberType	N	Nvarchar(50)			
	Address	N	Nvarchar(50)			

	AddressType	N	Nvarchar(50)			
	PostalCode	N	Nvarchar(50)			
	EmailAddress	N	Nvarchar(50)			
	Country	N	Nvarchar(50)			
	InsertDate	Y	datetime			SysDateTime
	ModifiedDate	Y	datetime			SysDateTime
	StartDate	Y	datetime			SysDateTime
	EndDate	Y	datetime			SysDateTime
DimCity	**CitySK**	Y	Int	Not Null	PK	Auto incrementing
	AlternateCityID	N	Int	Not Null		
	CityCode	N	Int			
	RegionKey	N	Int		FK	
	SrcModifiedDate	N	DateTime			
	InsertDate	Y	datetime			SysDateTime
	ModifiedDate	Y	datetime			SysDateTime
DimRegion	**RegionSK**	Y	Int	Not Null	PK	Auto incrementing
	AlternateRegionID	N	Int	Not Null		
	RegionCode	N	Int			
	SrcModifiedDate	N	DateTime			
	InsertDate	Y	datetime			SysDateTime
	ModifiedDate	Y	datetime			SysDateTime
DimCenter	**CenterSK**	Y	Int	Not Null	PK	Auto incrementing
	AlternateCenterID	N	Int	Not Null		
	CityKey	N	Int		FK	
	CenterManagerKey	N	Int		FK	
	CenterType	N	Nvarchar(50)			
	OpperationaArea	N	Nvarchar(50)			
	PhoneNumber	N	Nvarchar(25)			
	PhoneNumberType	N	Nvarchar(50)			
	NumberOfEmployees	N	Int			
	InsertDate	Y	datetime			SysDateTime

	ModifiedDate	Y	datetime			SysDateTime
	StartDate	Y	datetime			SysDateTime
	EndDate	Y	datetime			SysDateTime
FactWeeklyDemand	**DealID**	N	Int	Not Null		
	DealDateKey	N	int		FK	
	CenterKey	N	Int		FK	
	MealKey	N	int		FK	
	CheckoutPrice	N	Money			
	BasePrice	N	Money			
	EmailerForPromotion	N	Int			
	HomepageFeatured	N	Int			
	NumOrders	N	Int			
	TotalCheckoutPrice	Y	Money			([CheckoutPrice]*[NumOrders])
	TotalBasePrice	Y	Money			([BasePrice]*[NumOrders])
	TotalDiscountPerCenterMeal	Y	Money			(([BasePrice]*[NumOrders]) - ([CheckoutPrice]*[NumOrders]))
	DiscountAmount	Y	Money			([BasePrice] – [CheckoutPrice])
	DiscountPercent	Y	Money			((([BasePrice] – [CheckoutPrice])/[BasePrice])*100)
	InsertDate	Y	datetime			SysDateTime
	ModifiedDate	Y	datetime			SysDateTime

ETL Development

Data Extraction from Source tables to staging tables

Staging of each table was made in order as given below

1. Extract Meal Cuisine Data to Staging
2. Extract Meal Beverage Data to Staging
3. Extract Meal Data to Staging
4. Extract Region Data to Staging
5. Extract City Data to Staging
6. Extract CenterManager Data to Staging
7. Extract CenterManagerDetails Data to Staging
8. Extract Center Data to Staging
9. Extract CenterDetails Data to Staging
10. Extract WeeklyDemand to Staging

Derived columns in the Fact table

TotalCheckoutPrice = ([CheckoutPrice]*[NumOrders])
TotalBasePrice = ([BasePrice]*[NumOrders])
TotalDiscountPerCenterMeal = (([BasePrice]*[NumOrders]) - ([CheckoutPrice]*[NumOrders]))
DiscountAmount = ([BasePrice] – [CheckoutPrice])
DiscountPercent = ((([BasePrice] – [CheckoutPrice])/[BasePrice])*100)

Staging package name – MealDemand_Load_Staging.dtsx

Durring the staging process all data from sources will be extracted and loaded into the **MealDemand_Staging** Database.

Following are the names of the table to which the data was loaded.

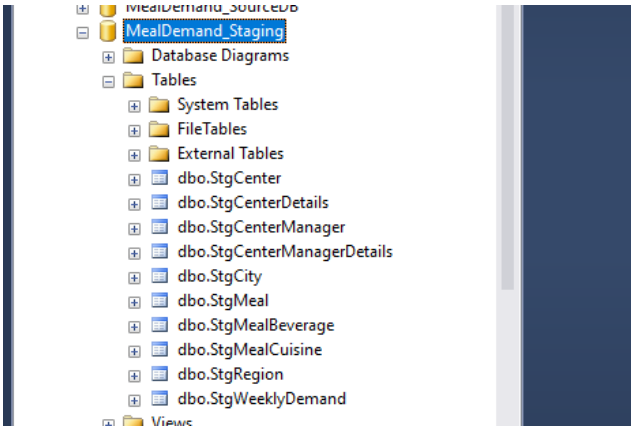


Figure 1: Staging Tables and Database

Staging of all tables.

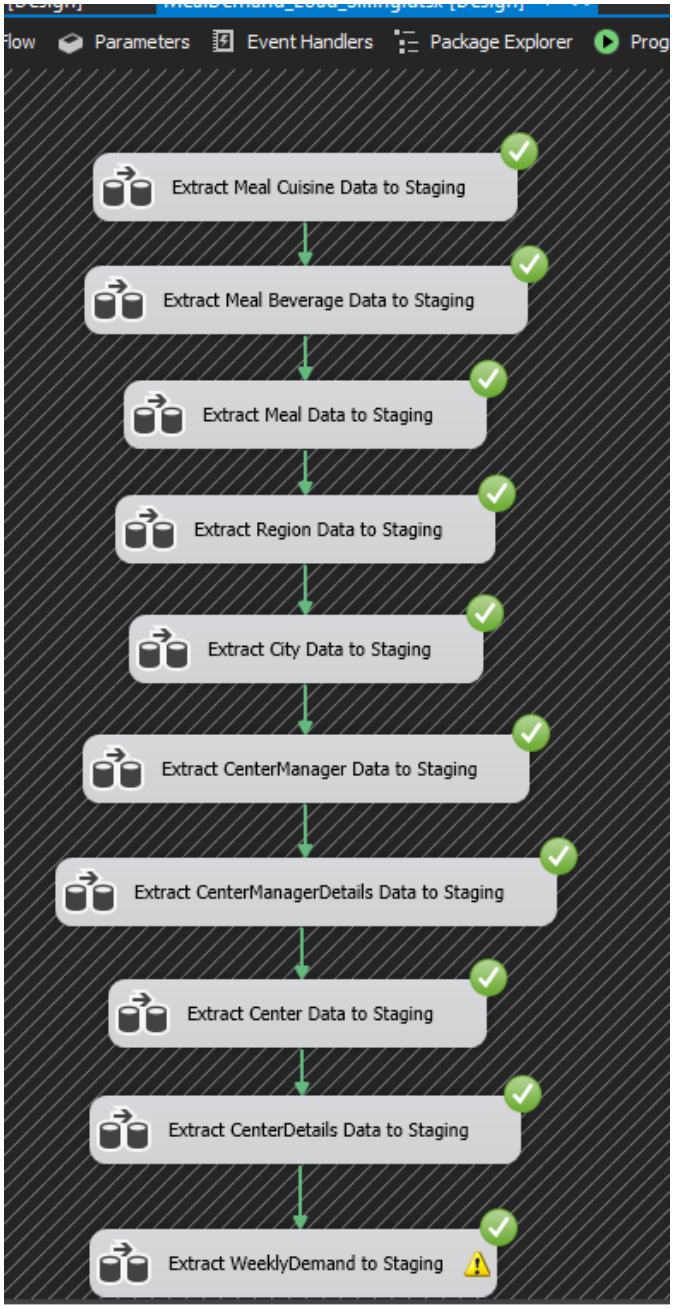


Figure 2: MealDemand_Load_Staging

1. Extract Meal Cuisine Data to Staging

Figure 3: Extract Meal Cuisine Data to Staging

2. Extract Meal Beverage Data to Staging

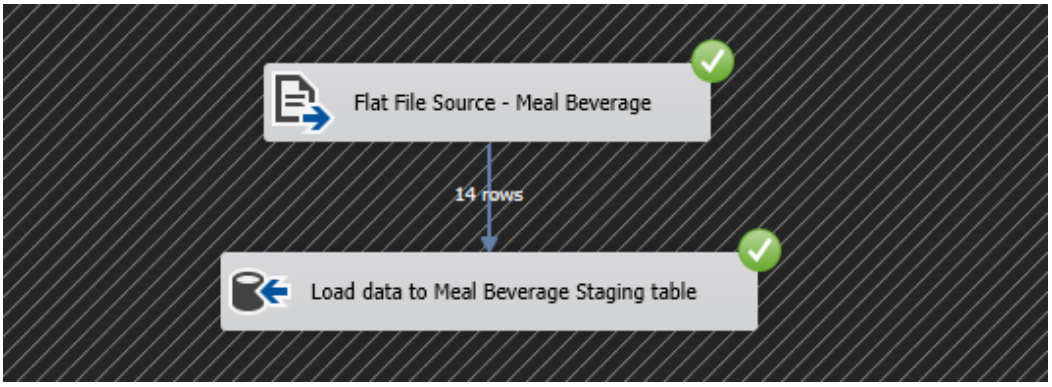


Figure 4:Extract Meal Beverage Data to Staging

3. Extract Meal Data to Staging

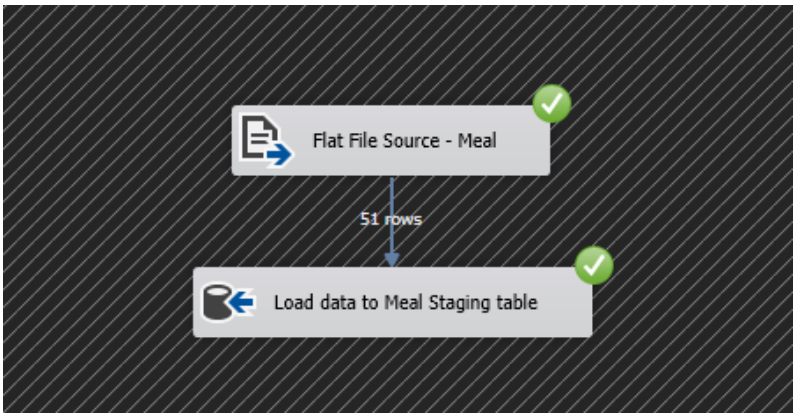


Figure 5: Extract Meal Data to Staging

4. Extract Region Data to Staging

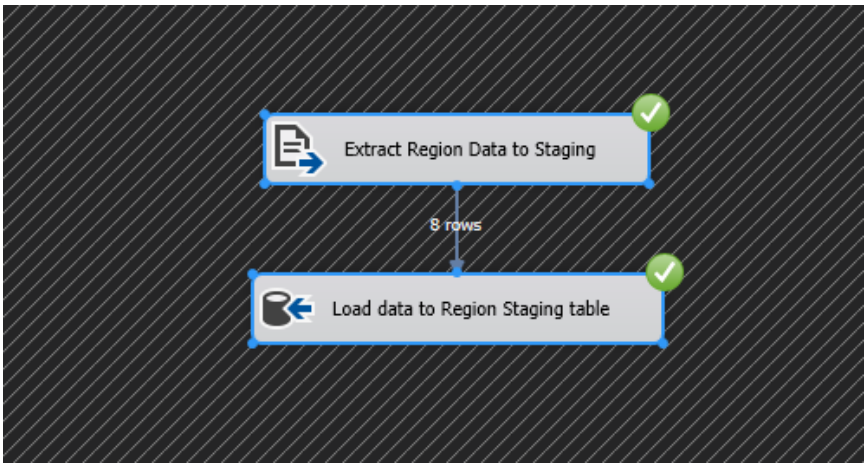


Figure 6: Extract Region Data to Staging

5. Extract City Data to Staging

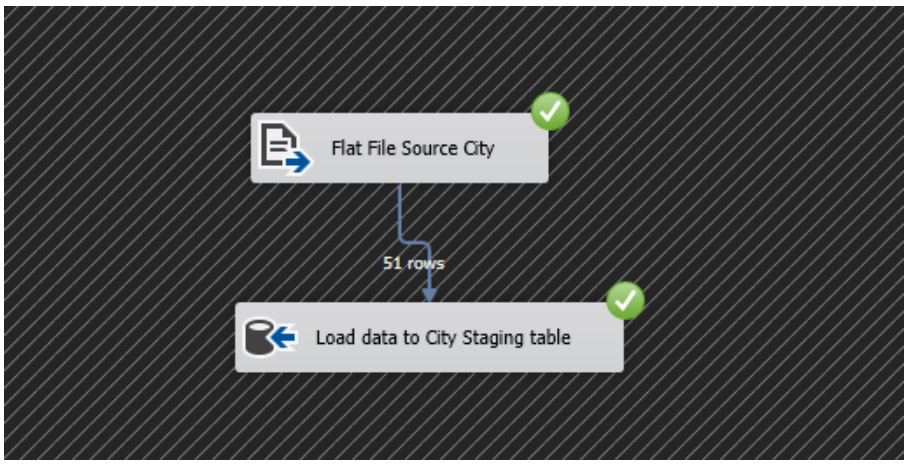


Figure 7: Extract City Data to Staging

6. Extract CenterManager Data to Staging

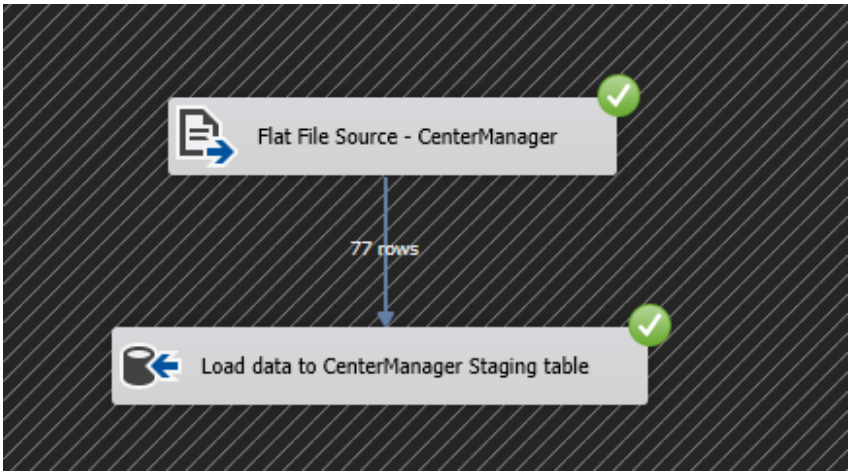


Figure 8: Extract CenterManager Data to Staging

7. Extract CenterManagerDetails Data to Staging

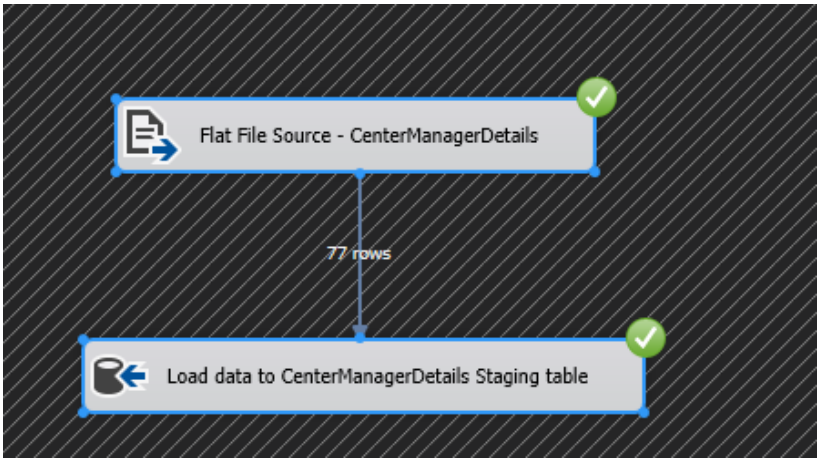


Figure 9: Extract CenterManagerDetails Data to Staging

8. Extract Center Data to Staging

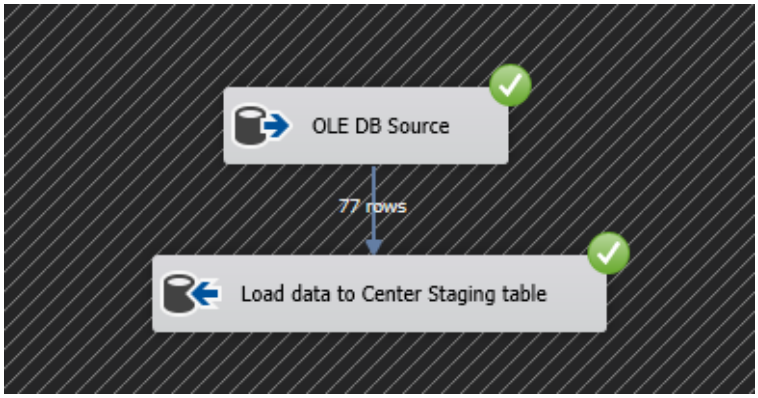


Figure 10: Extract Center Data to Staging

9. Extract CenterDetails Data to Staging

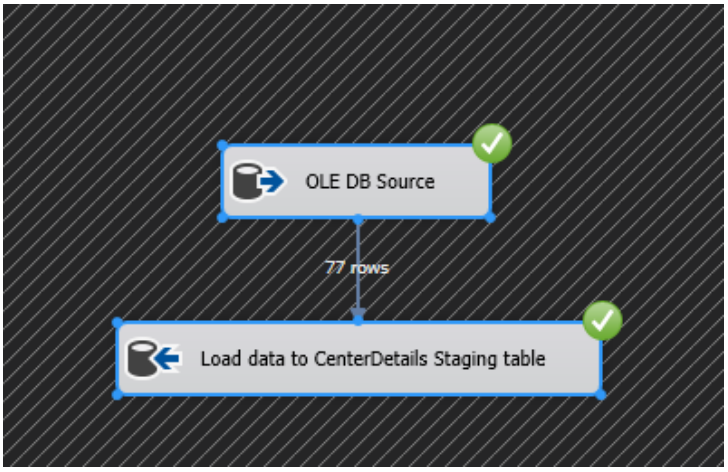


Figure 11: Extract CenterDetails Data to Staging

10. Extract WeeklyDemand to Staging

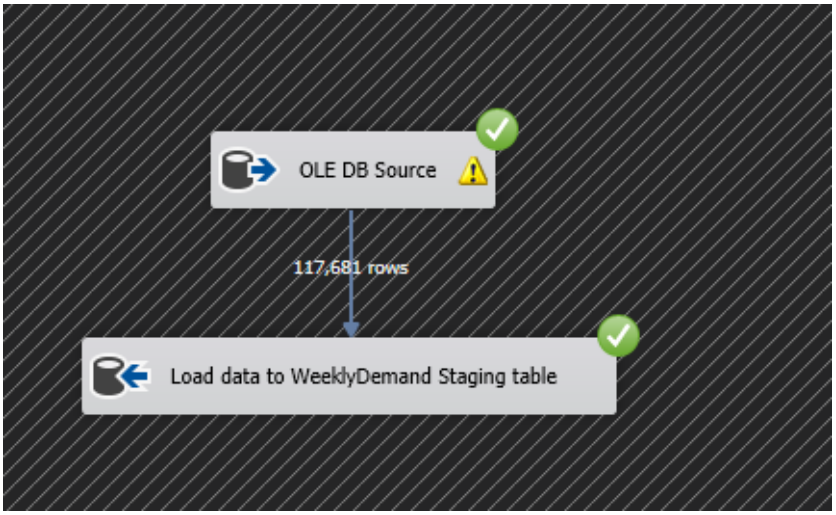
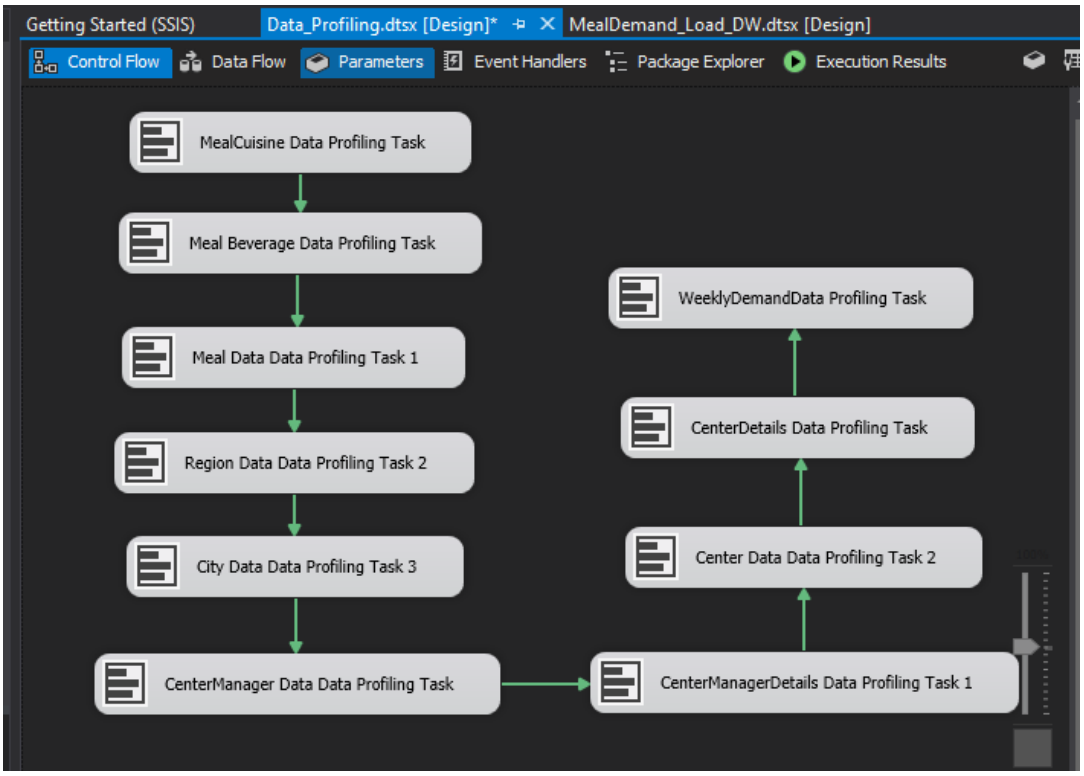


Figure 12: Extract WeeklyDemand to Staging

Data Profiling



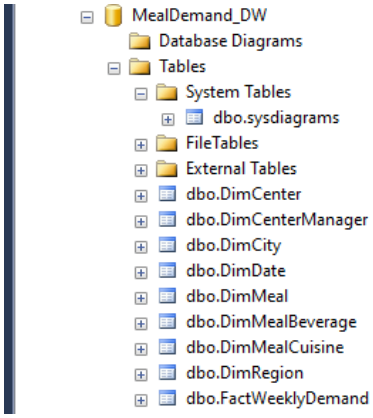
Transform and Load to Data Ware House

Dimension tables were loaded to the data ware house based on the following order

1. DimMealCuisine - Transform and Load MealCuisine Data to Data Warehouse
2. DimMealBeverage - Transform and Load MealBeverage Data to Data Warehouse
3. DimMeal - Transform and Load Meal Data to Data Warehouse
4. DimCenterManager - Transform and Load CenterManager Data to Data Warehouse
5. DimRegion - Transform and Load Region Data to Data Warehouse
6. DimCity - Transform and Load City Data to Data Warehouse
7. DimCenter - Transform and Load Center Data to Data Warehouse
8. WeeklyDemand Fact - Transform and WeeklyDemand Fact Table to Data Warehouse

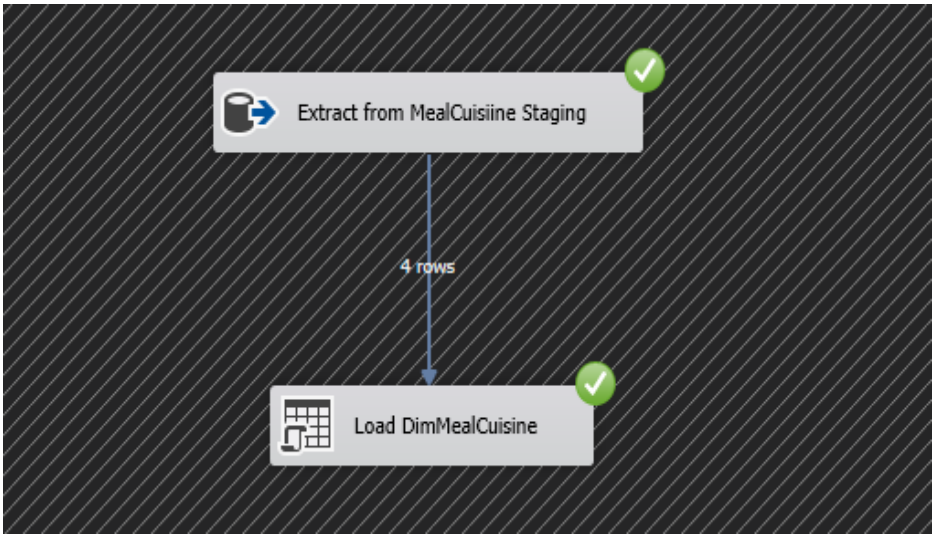
Slowly Changing Dimension - DimCenterManager
Slowly Changing Dimension – DimCenter

Durring the process of loading the Staging layer data will be loaded to the **MealDemand_DW** database

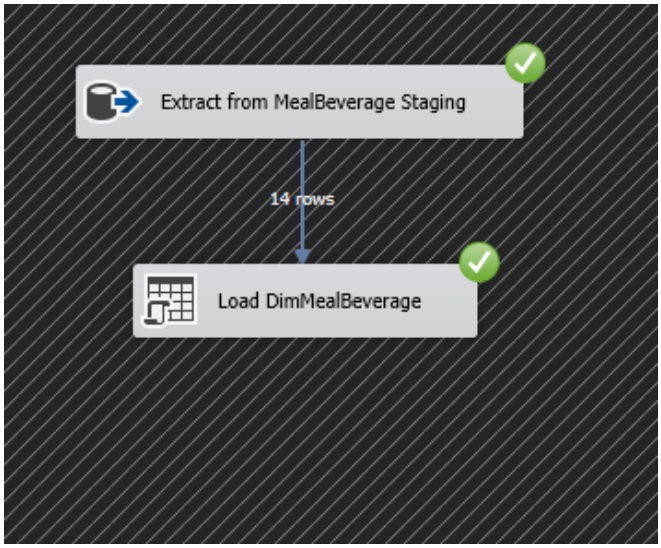


Name of package: **MealDemand_Load_DW.dtsx**

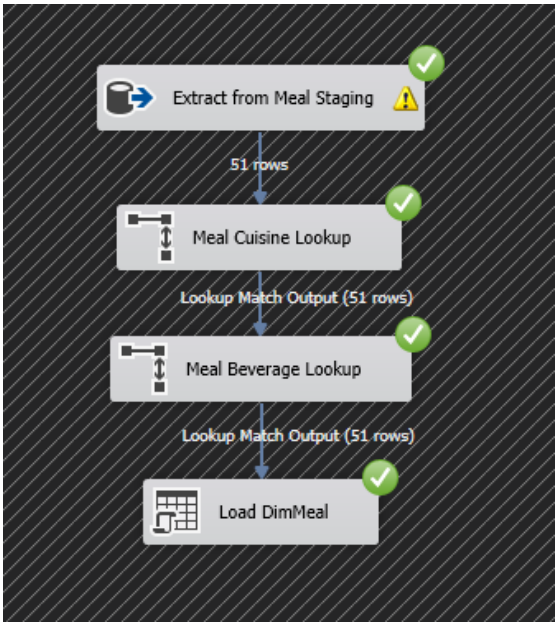
1. DimMealCuisine - Transform and Load MealCuisine Data



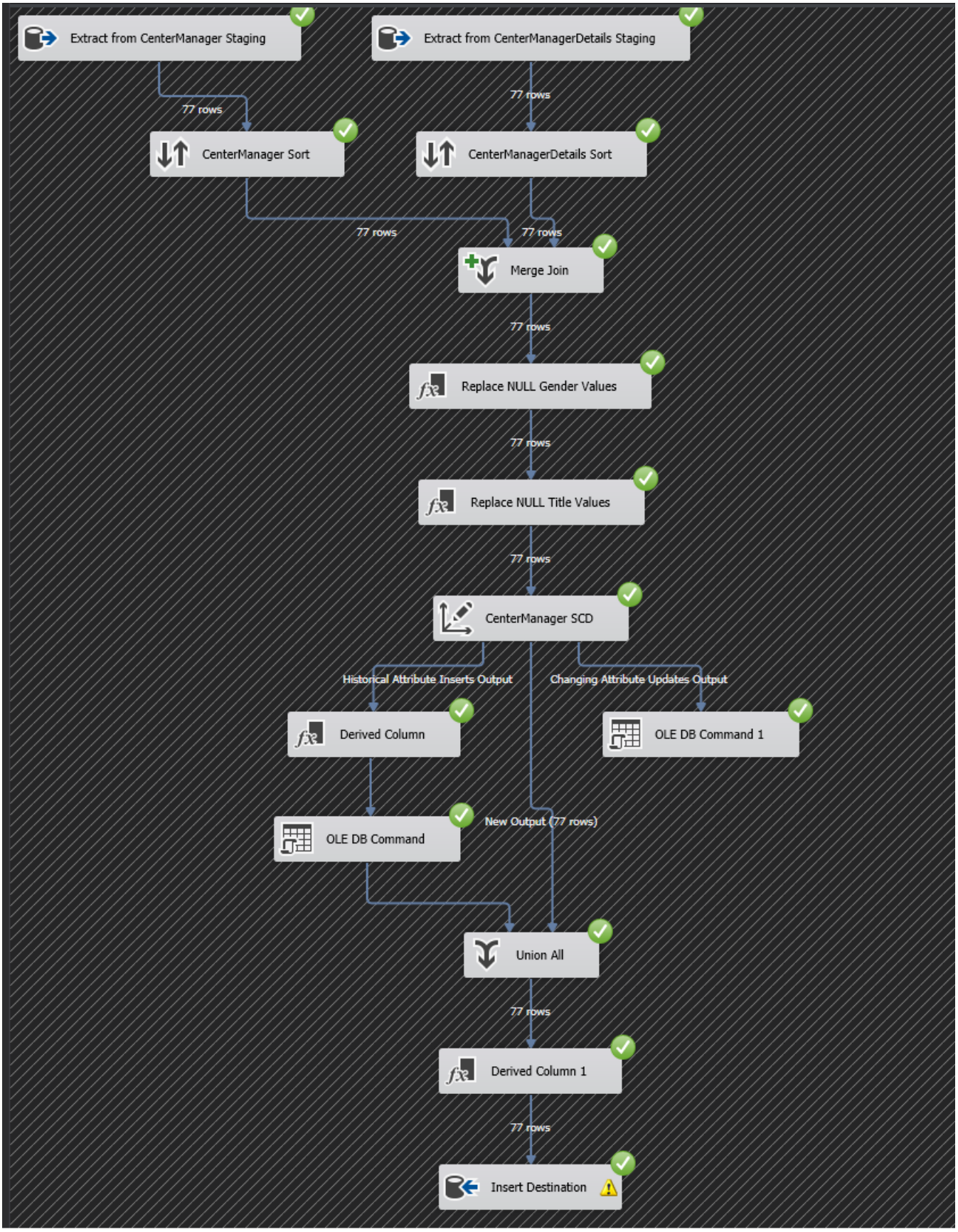
2. DimMealBeverage - Transform and Load MealBeverage Data



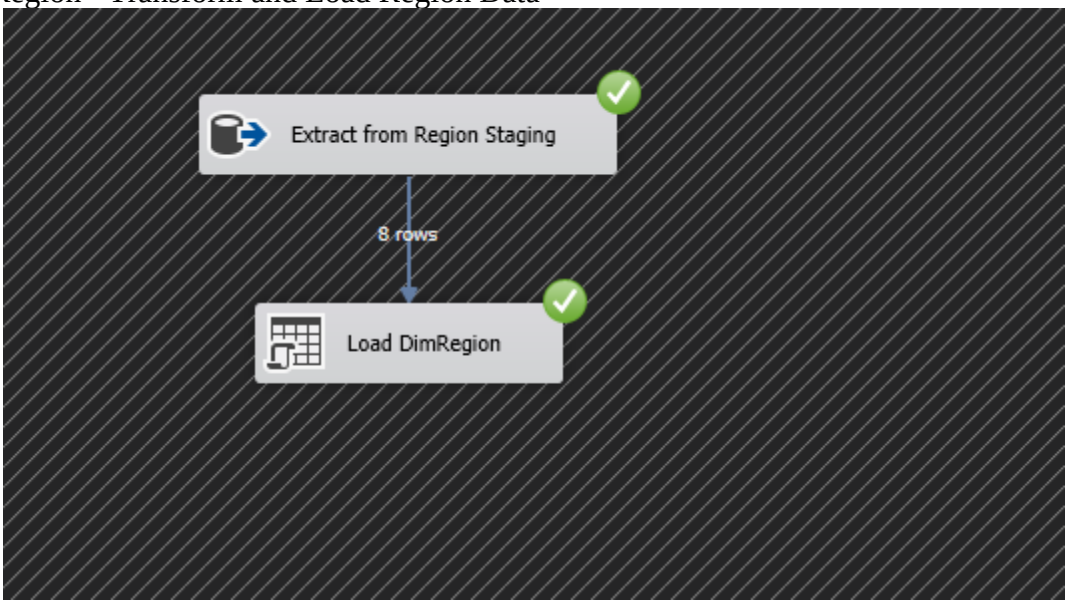
3. DimMeal - Transform and Load Meal Data



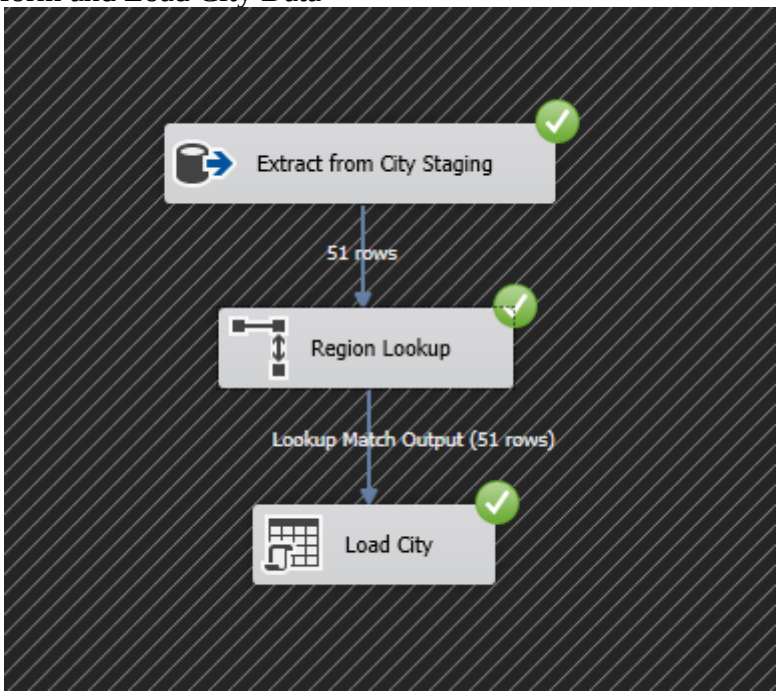
4. DimCenterManager - Transform and Load CenterManager Data


Extract from CenterManager Staging
Extract from CenterManagerDetails Staging
77 rows
77 rows
CenterManager Sort
CenterManagerDetails Sort
77 rows
77 rows
Merge Join
77 rows
Replace NULL Gender Values
77 rows
Replace NULL Title Values
77 rows
CenterManager SCD
Historical Attribute Inserts Output
Changing Attribute Updates Output
Derived Column
OLE DB Command 1
OLE DB Command
New Output (77 rows)
Union All
77 rows
Derived Column 1
77 rows
Insert Destination

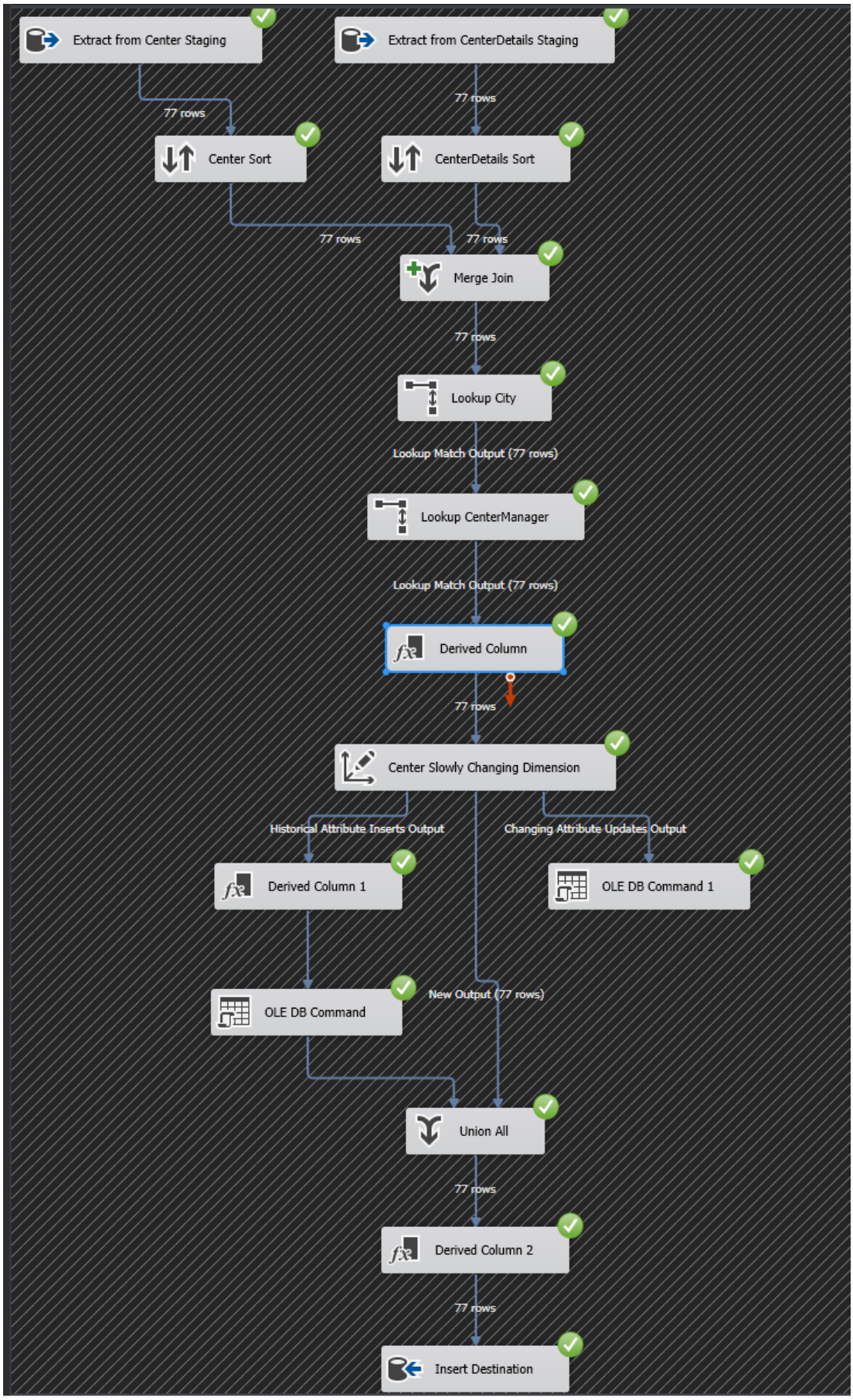
5. DimRegion - Transform and Load Region Data



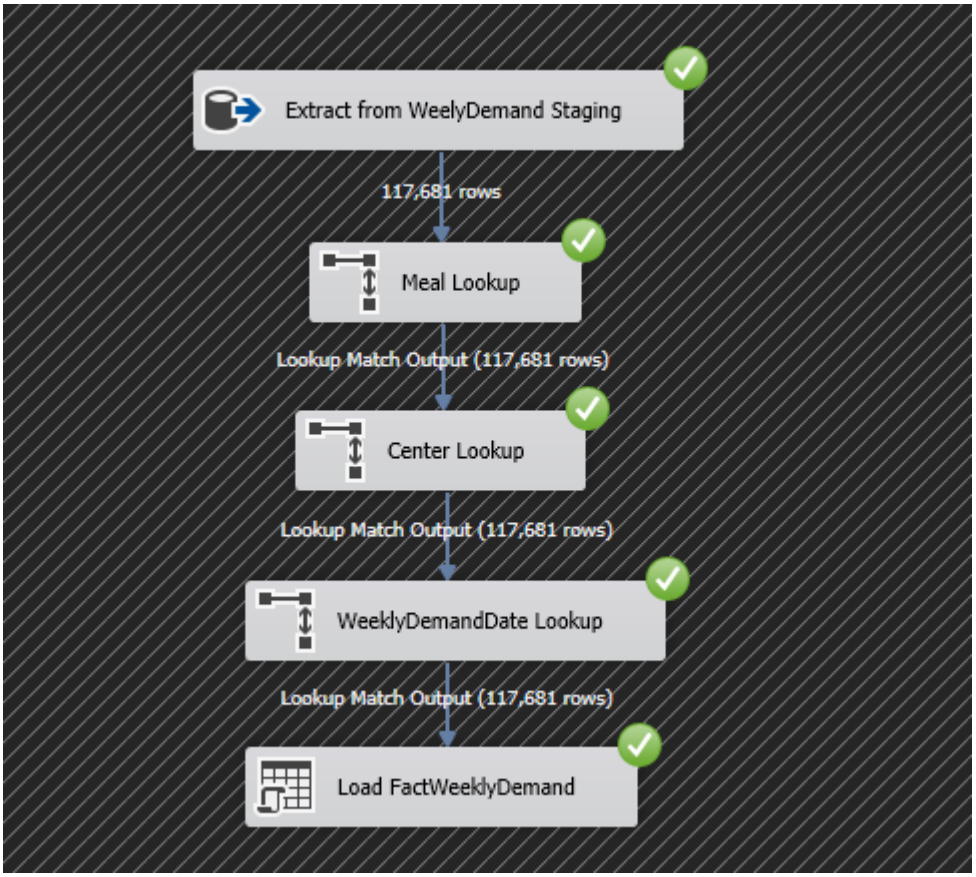
6. DimCity - Transform and Load City Data



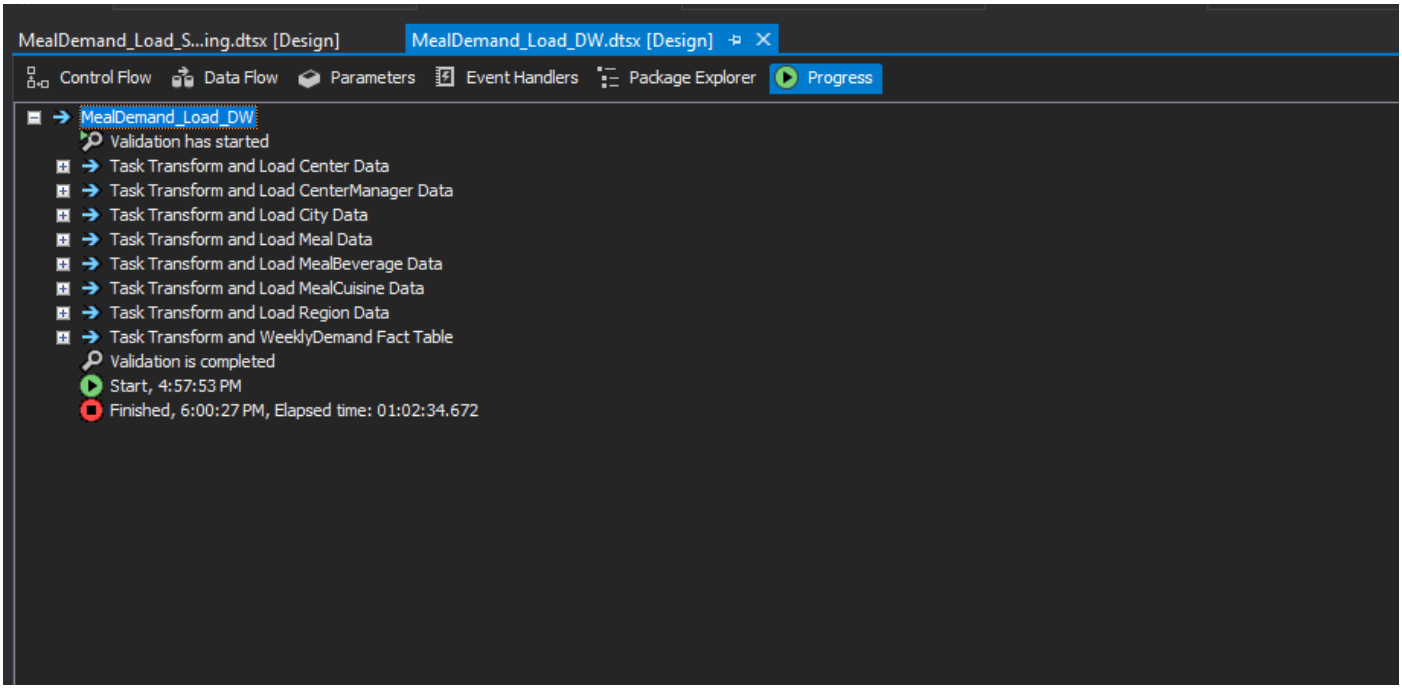
7. DimCenter - Transform and Load Center Data



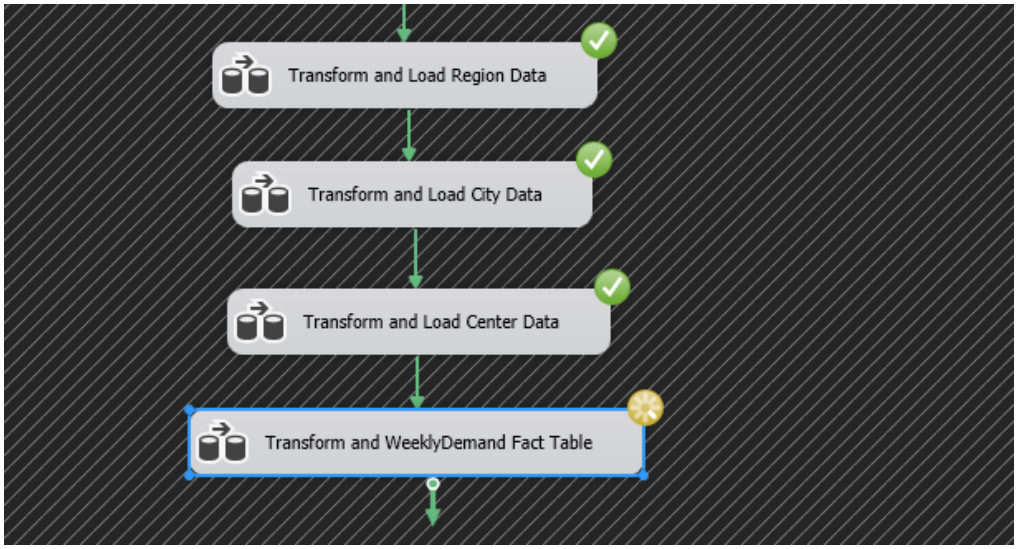
8. WeeklyDemand **Fact** - Transform and WeeklyDemand Fact Table



Meal Demand Load Progress




Transform and Load MealCuisine Data
Transform and Load MealBeverage Data
Transform and Load Meal Data
Transform and Load CenterManager Data
Transform and Load Region Data
Transform and Load City Data


Transform and Load Region Data
Transform and Load City Data
Transform and Load Center Data
Transform and WeeklyDemand Fact Table