Sri Lanka Institute of Information Technology

Data Warehousing & Business Intelligence

Assignment 1

2022

M.A.D.G.A. SURIYAWATTA

IT20135652

TABLE OF CONTENTS

STEP 1: DATA SET SELECTION

This data set about investigation data for aviation accidents and incidents from 2002 to 2007. An occurrence associated with the operation of an aircraft, which takes place from the time any person boards the aircraft with the intention of flight until all such persons have disembarked, and in which a) a person is fatally or seriously injured, b) the aircraft sustains significant damage or structural failure, or c) the aircraft goes missing or becomes completely inaccessible. This data set defines an aviation incident as an occurrence, other than an accident, associated with the operation of an aircraft that affects or could affect the safety of operation. Accidents and incidents are investigated by government bodies such as the FAA and NTSB.

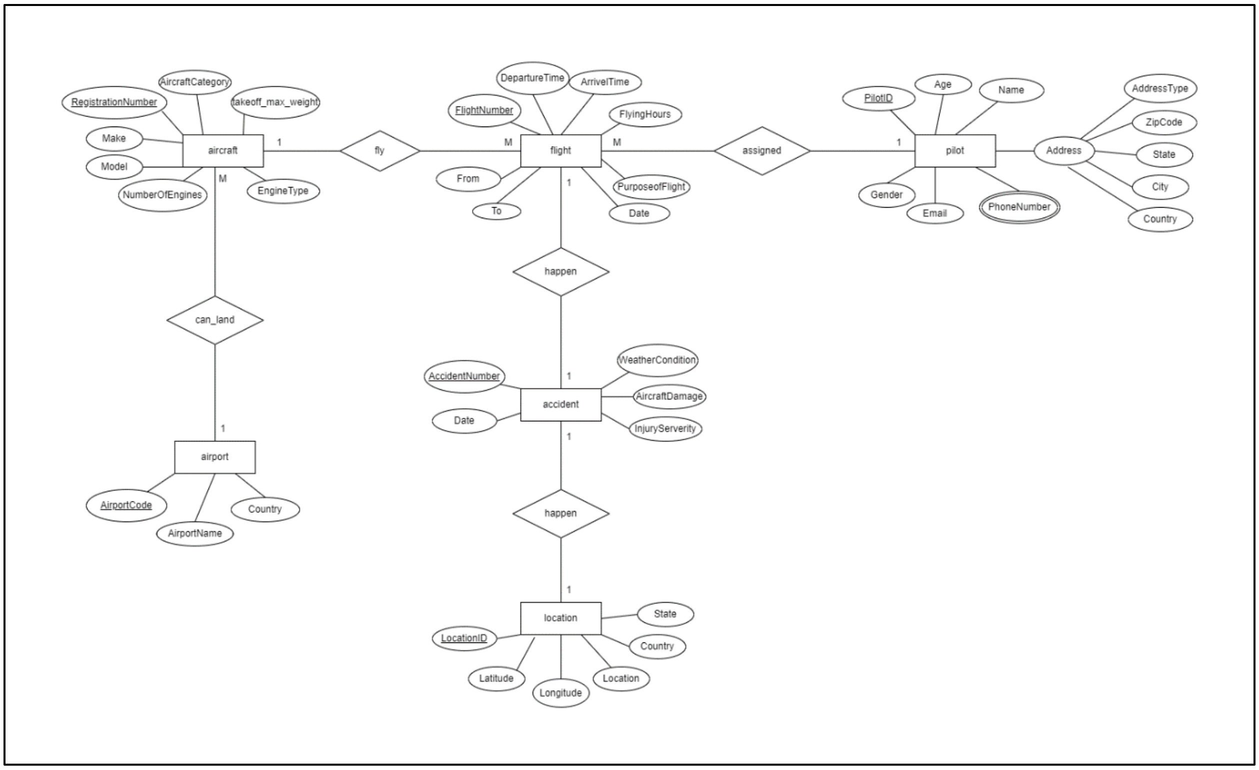
Link to the selected source data set:

https://www.kaggle.com/datasets/prathamsharma123/aviation-accidents-and-incidents-ntsb-faa-waas

The original dataset has less tables. I cut the columns of original source tables and put them into different source tables to get more dimensions and a hierarchy, because the assignment document says that we need to enrich the ETL process.

Customized source has seven tables and it include accidents' details, locations' details, aircrafts' details, airports' details, pilot's details, pilots' addresses' details.

The ER Diagram of the Data Set



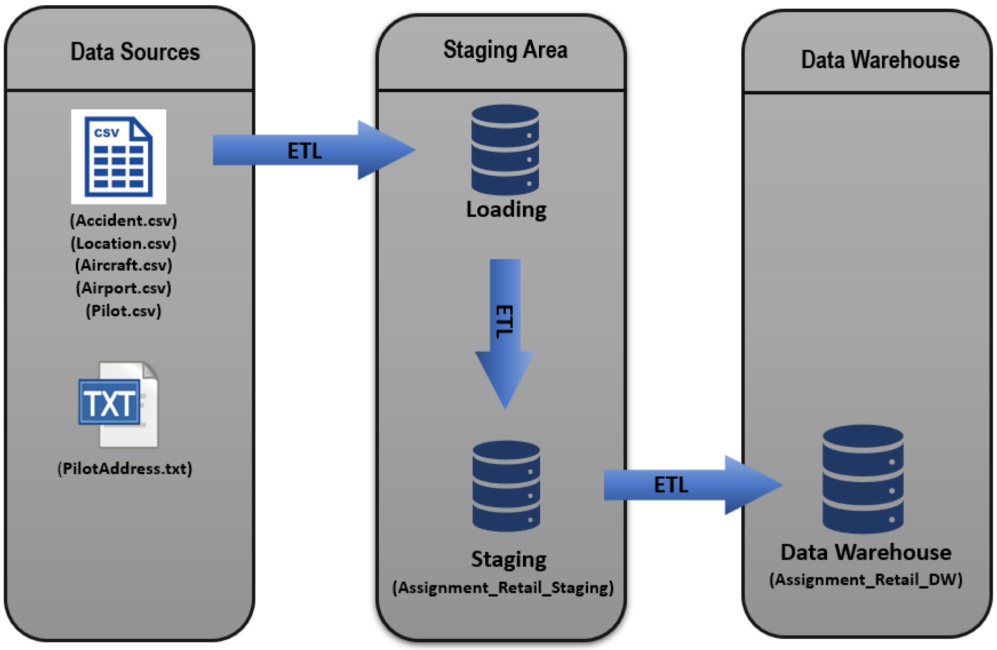
STEP 2: PREPARATION OF DATA SOURCES

There are seven source tables in two formats (.csv & .txt). And they were used to create the following,

Data Source Name	Data Source Type	Description
Accident	CSV	Details about accidents
Location	CSV	Details about accidents happened locations
Flight	CSV	Details about the flights involved the accidents
Aircraft	CSV	Details about the aircrafts involved the accidents
Airport	CSV	Details about the airport where the aircrafts land
Pilot	CSV	Details about the pilots who involved the flights

Pilot Addresses	Text	Details abouts pilots' addresses	

STEP 3: SOLUTION ARCHITECTURE



Above architecture shows the high-level BI solution to the warehouse design.

<u>Data Sources</u>

csv' component is used to display Comma Separated files and '.txt' component represents Text files.

<u>Staging Area</u>

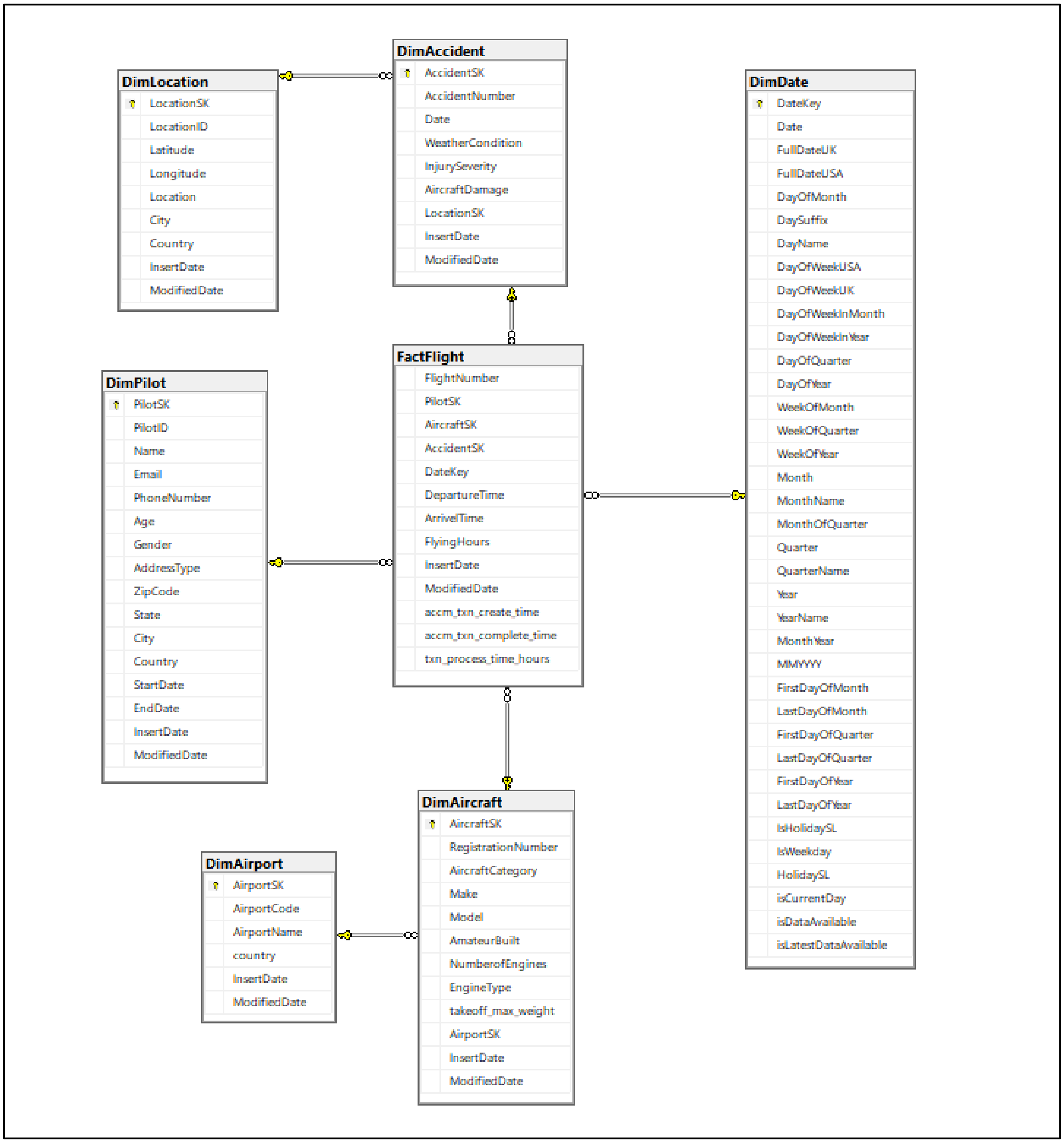
This area Loading DB component represents the process of the creating database tables. Accident, Location, Aircraft, Airport, Flight, Pilot and Pilot Addresses text files was imported to the database and was used to create the tables. And these tables were used as the DB source data. Staging DB component represents creating staging level tables through the 'Extract'.

<u>Data Warehouse</u>

Data warehouse DB component is used display the cratering dimension tables in the warehouse using ‘Transform’ and ‘Load.’

STEP 4: DATA WAREHOUSE DESIGN & DEVELOPMENT

Following figure will show how the fact table and dimension tables was combined in a rational manner. For this scenario, **snowflake schema** type was used.



DimAccident

AccidentSK
AccidentNumber
Date
WeatherCondition
InjurySeverity
AircraftDamage
LocationSK
InsertDate
ModifiedDate

DimLocation

LocationSK
LocationID
Latitude
Longitude
Location
City
Country
InsertDate
ModifiedDate

DimDate

DateKey
Date
FullDateUK
FullDateUSA
DayOfMonth
DaySuffix
DayName
DayOfWeekUSA
DayOfWeekUK
DayOfWeekInMonth
DayOfWeekInYear
DayOfQuarter
DayOfYear
WeekOfMonth
WeekOfQuarter
WeekOfYear
Month
MonthName
MonthOfQuarter
Quarter
QuarterName
Year
YearName
MonthYear
MMYYYY
FirstDayOfMonth
LastDayOfMonth
FirstDayOfQuarter
LastDayOfQuarter
FirstDayOfYear
LastDayOfYear
IsHolidaySL
IsWeekday
HolidaySL
isCurrentDay
isDataAvailable
isLatestDataAvailable

FactFlight

FlightNumber
PilotSK
AircraftSK
AccidentSK
DateKey
DepartureTime
ArrivelTime
FlyingHours
InsertDate
ModifiedDate
accm_txn_create_time
accm_txn_complete_time
txn_process_time_hours

DimPilot

PilotSK
PilotID
Name
Email
PhoneNumber
Age
Gender
AddressType
ZipCode
State
City
Country
StartDate
EndDate
InsertDate
ModifiedDate

DimAircraft

AircraftSK
RegistrationNumber
AircraftCategory
Make
Model
AmateurBuilt
NumberofEngines
EngineType
takeoff_max_weight
AirportSK
InsertDate
ModifiedDate

DimAirport

AirportSK
AirportCode
AirportName
country
InsertDate
ModifiedDate

Dimension Types

Hierarchical Dimension

- Date – all the hierarchies in date
- Pilot Addresses – Country -> City -> State -> ZipCode
- Location - Location -> City -> Country

Slowly Changing Dimension

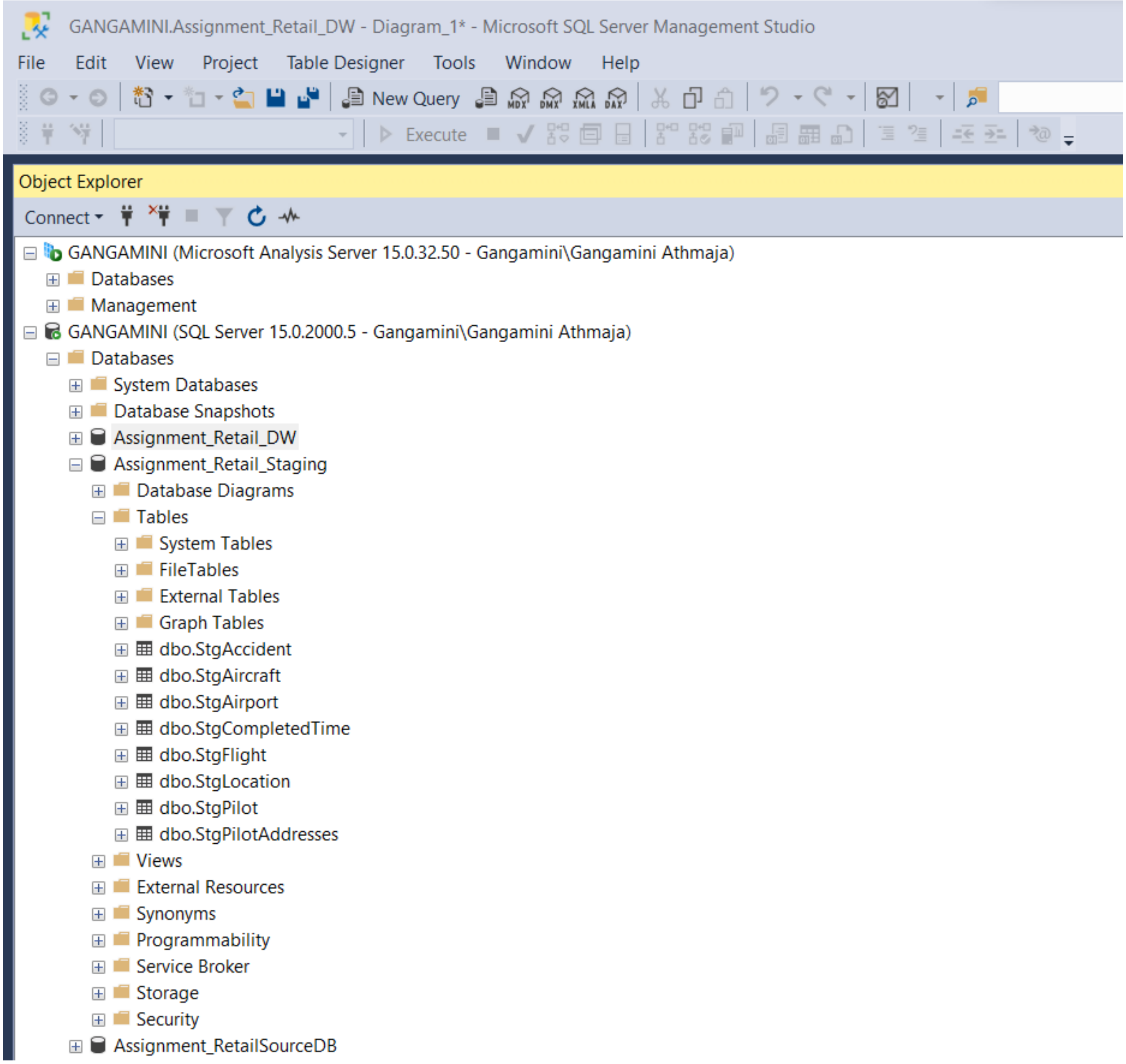
DimPilot is slowly changing dimention. PhoneNumber may be changed in future. Therefore, I get it as slowly changing Attribute.

STEP 5: ETL DEVELOPMENT

1.Extract

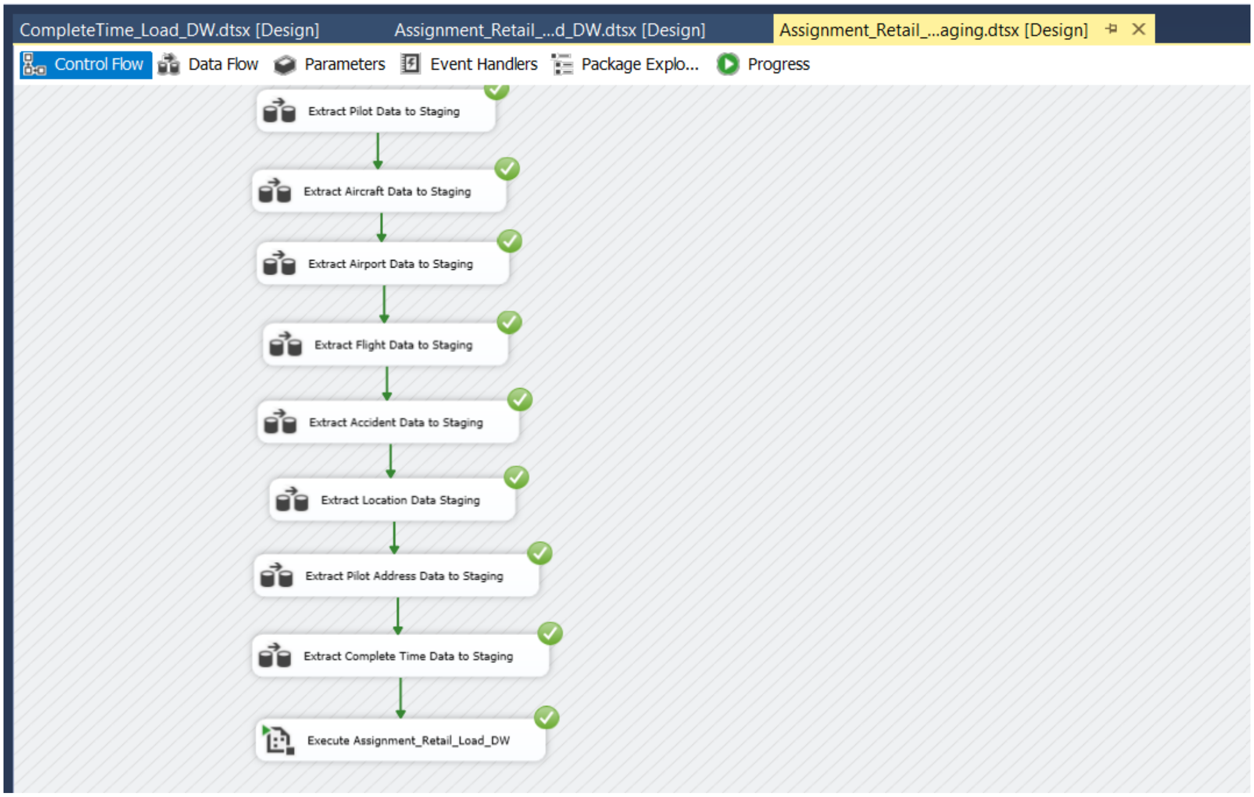
In this step, All the data sources were imported to the staging tables by using the relevant Data connection. Flat file connection was used for text files and csv files. All those tables were imported to the Assignment_Retail_Satging.

- **Snapshot of SSMS Staging Database**



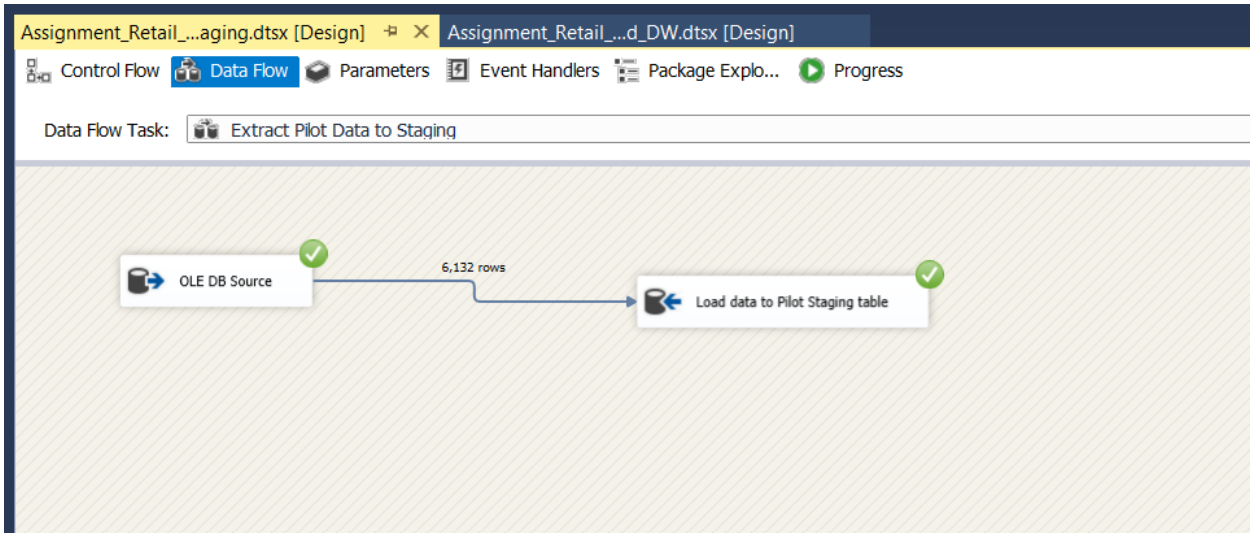
- GANGAMINI (Microsoft Analysis Server 15.0.32.50 - Gangamini\Gangamini Athmaja)
  - Databases
  - Management
- GANGAMINI (SQL Server 15.0.2000.5 - Gangamini\Gangamini Athmaja)
  - Databases
    - System Databases
    - Database Snapshots
    - Assignment_Retail_DW
    - Assignment_Retail_Staging
      - Database Diagrams
      - Tables
        - System Tables
        - FileTables
        - External Tables
        - Graph Tables
        - dbo.StgAccident
        - dbo.StgAircraft
        - dbo.StgAirport
        - dbo.StgCompletedTime
        - dbo.StgFlight
        - dbo.StgLocation
        - dbo.StgPilot
        - dbo.StgPilotAddresses
      - Views
      - External Resources
      - Synonyms
      - Programmability
      - Service Broker
      - Storage
      - Security
    - Assignment_RetailSourceDB

- **Snapshot of Visual Studio Control Flow of Extract**


CompleteTime_Load_DW.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Assignment_Retail_...aging.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Extract Pilot Data to Staging
Extract Aircraft Data to Staging
Extract Airport Data to Staging
Extract Flight Data to Staging
Extract Accident Data to Staging
Extract Location Data Staging
Extract Pilot Address Data to Staging
Extract Complete Time Data to Staging
Execute Assignment_Retail_Load_DW

- **Snapshots of several data types of Data Flows**

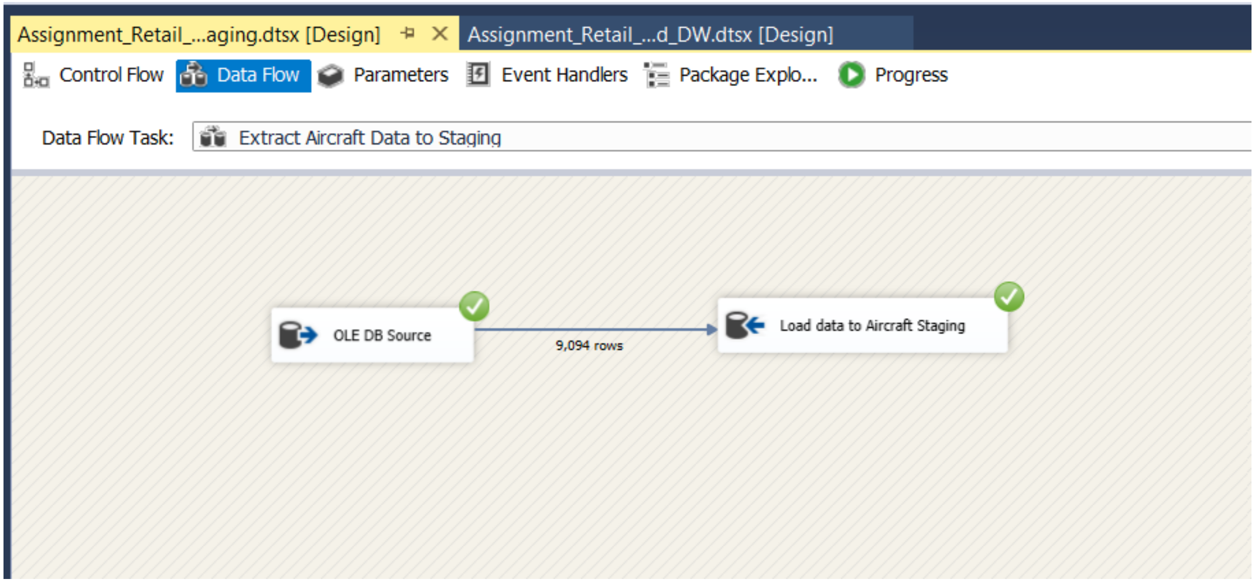
- Extract Pilot Data to staging


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress

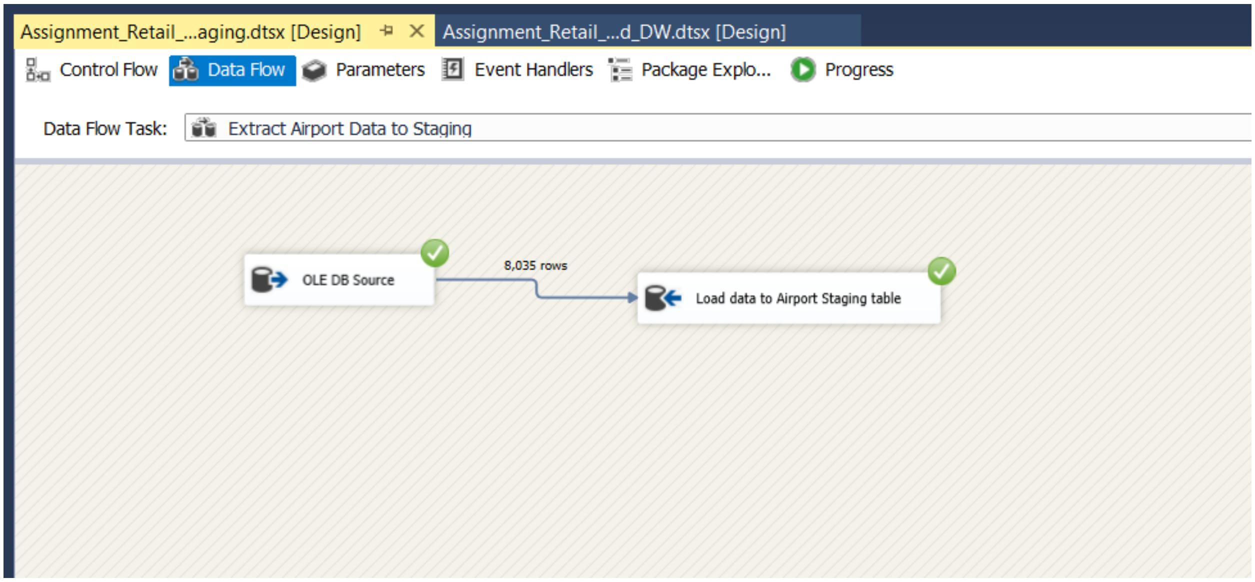
Data Flow Task: Extract Pilot Data to Staging

OLE DB Source
6,132 rows
Load data to Pilot Staging table

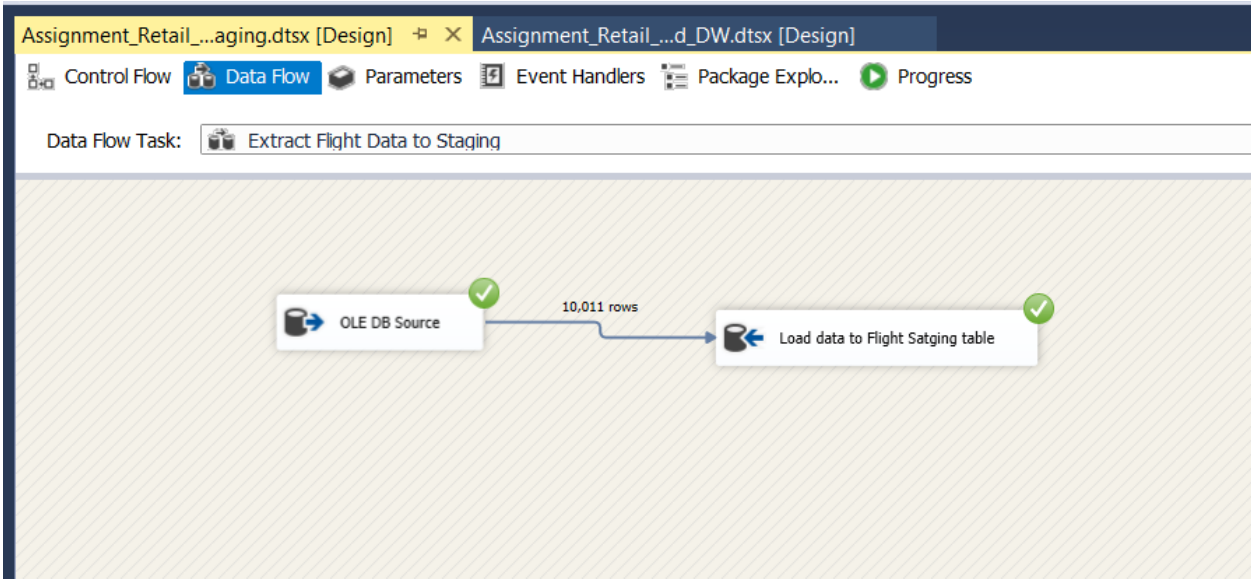
- Extract Aircraft Data to Staging


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Data Flow Task:
Extract Aircraft Data to Staging
OLE DB Source
9,094 rows
Load data to Aircraft Staging

- Extract Airport Data to staging


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Data Flow Task:
Extract Airport Data to Staging
OLE DB Source
8,035 rows
Load data to Airport Staging table

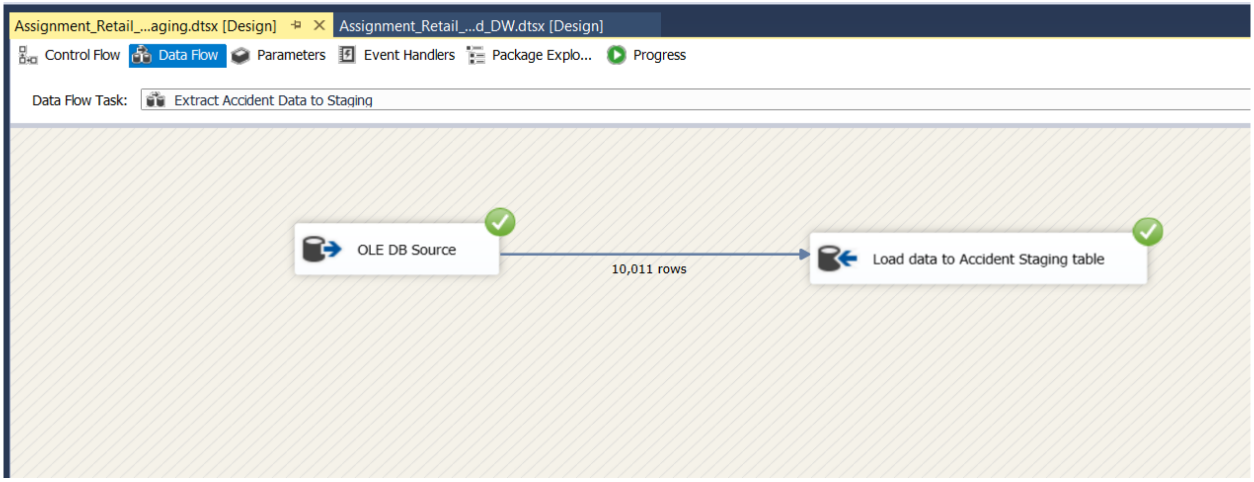
- Extract Flight Data to Staging



Data Flow Task: Extract Flight Data to Staging

Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
OLE DB Source
10,011 rows
Load data to Flight Satging table

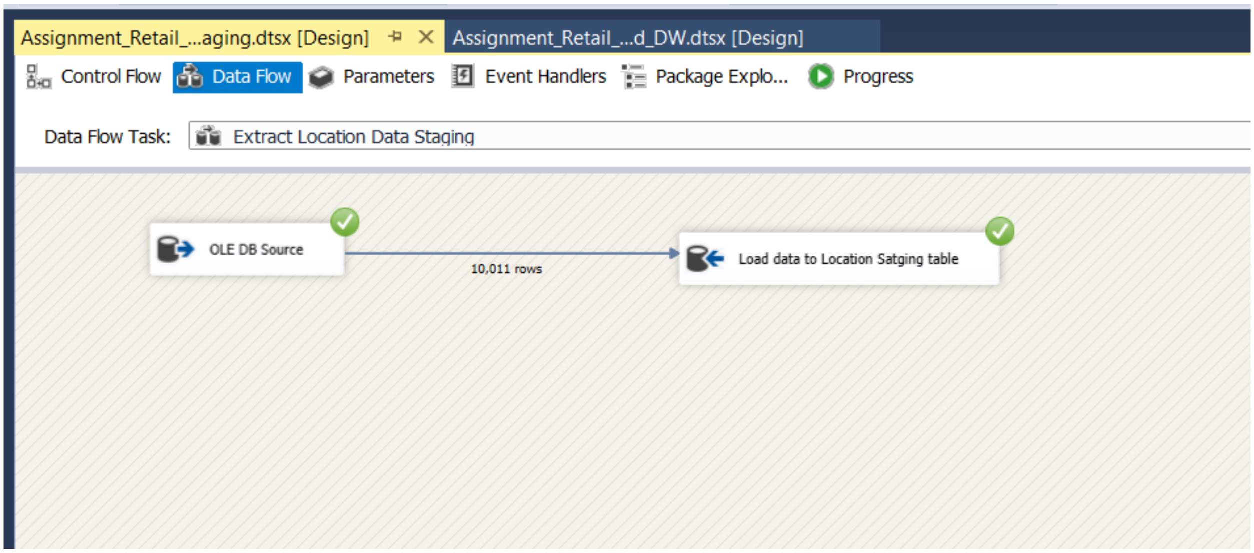
- Extract Accident Data to Staging



Data Flow Task: Extract Accident Data to Staging

Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
OLE DB Source
10,011 rows
Load data to Accident Staging table

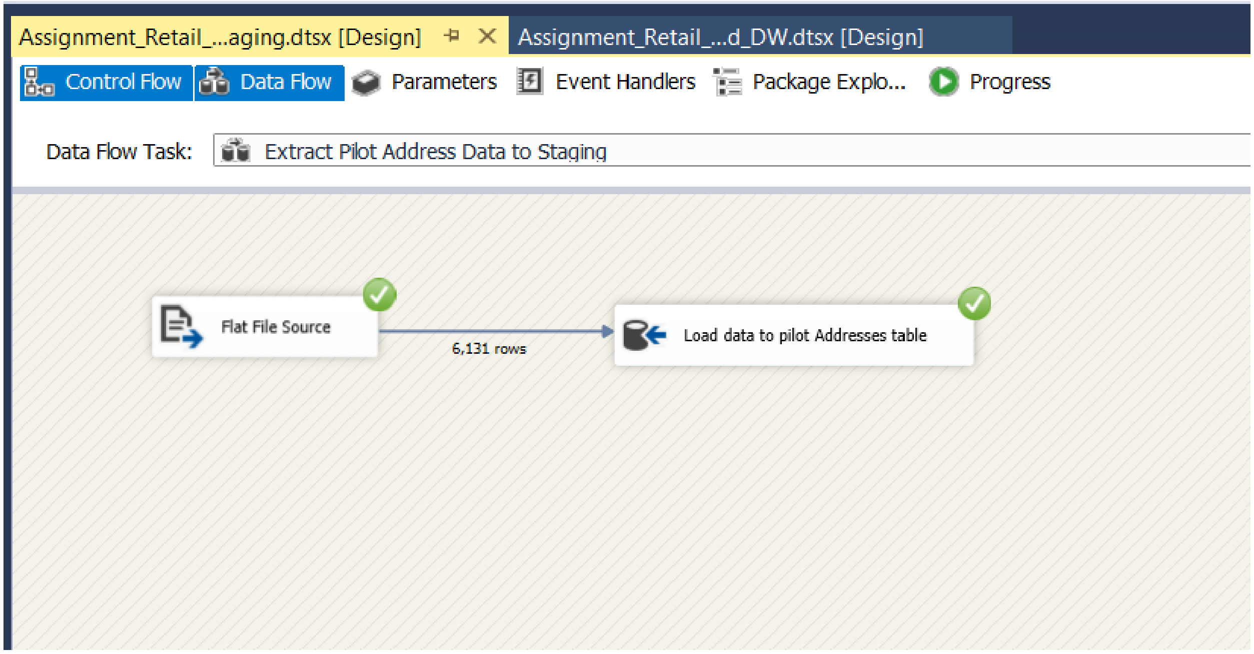
- Extract Location Data to Staging



Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
OLE DB Source
10,011 rows
Load data to Location Satging table

Data Flow Task: Extract Location Data Staging

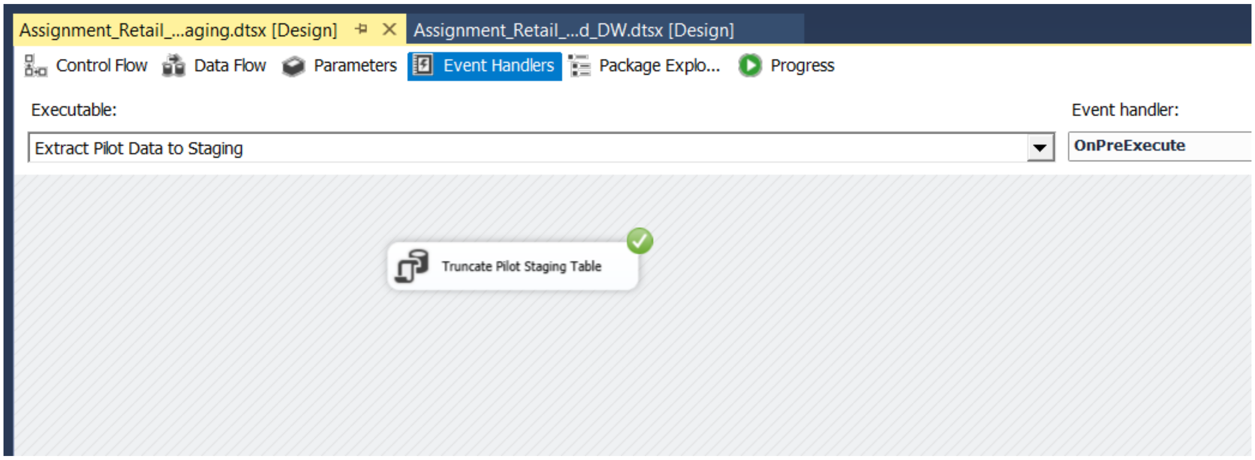
- Extract Pilot Addresses Data to Staging



Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Flat File Source
6,131 rows
Load data to pilot Addresses table

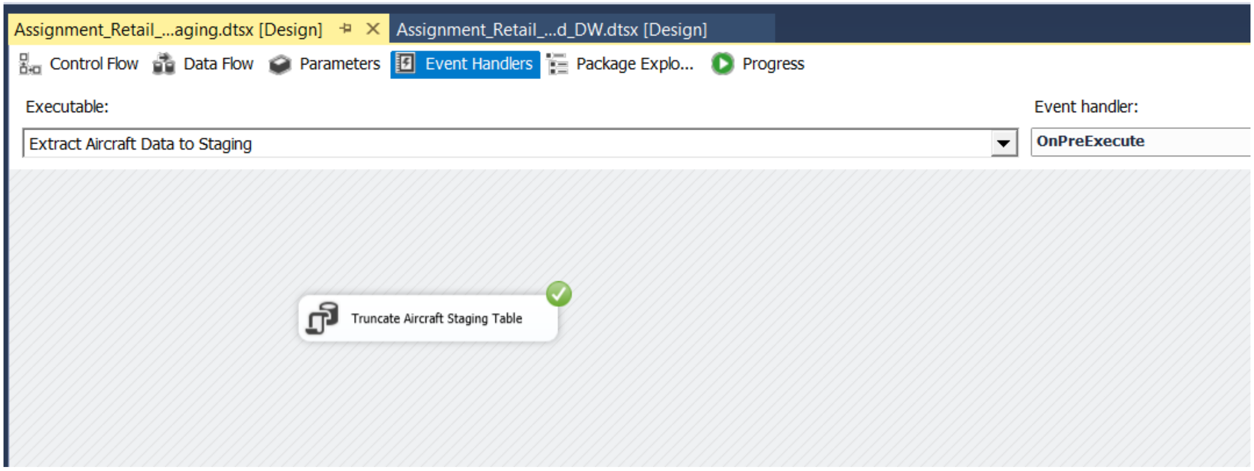
Data Flow Task: Extract Pilot Address Data to Staging

2. Event Handling (Truncate Staging Data)


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Truncate Pilot Staging Table

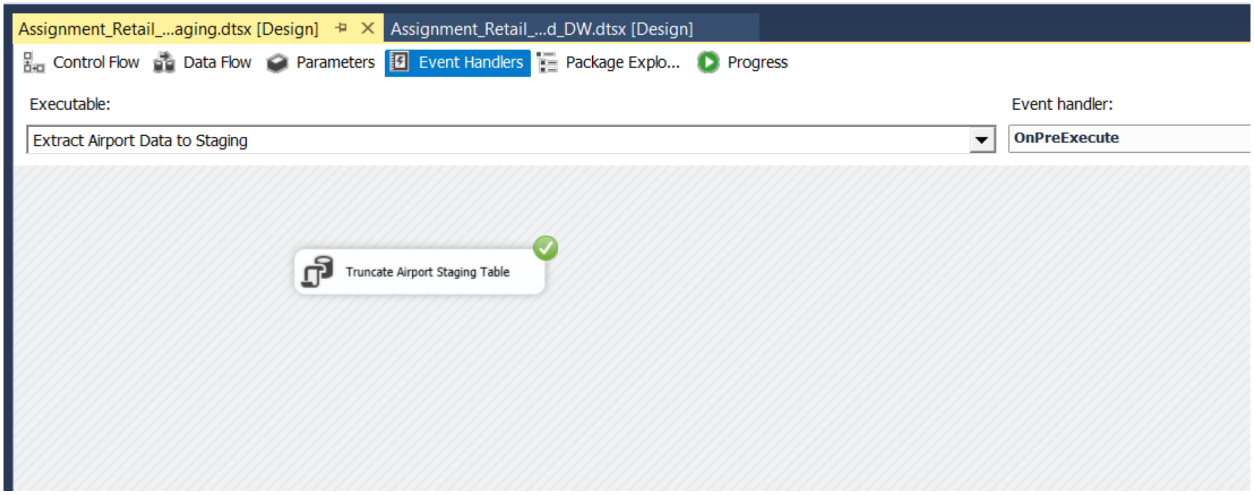
Executable: Extract Pilot Data to Staging

Event handler: OnPreExecute


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Truncate Aircraft Staging Table

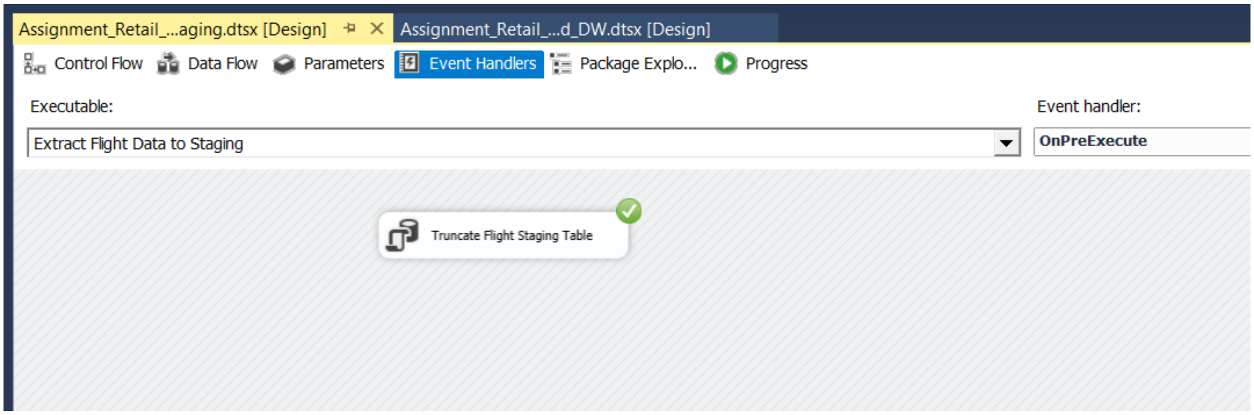
Executable: Extract Aircraft Data to Staging

Event handler: OnPreExecute


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Truncate Airport Staging Table

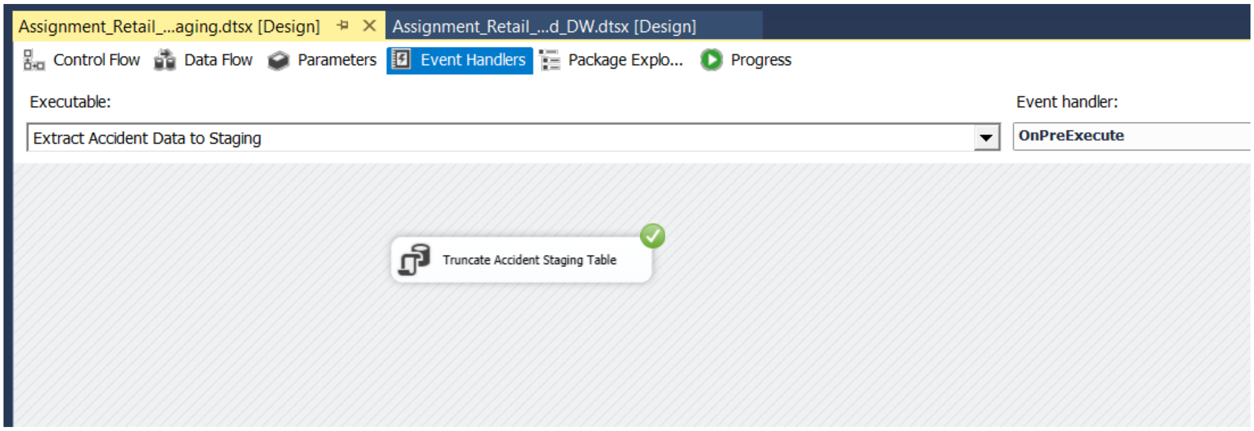
Executable: Extract Airport Data to Staging

Event handler: OnPreExecute


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Truncate Flight Staging Table

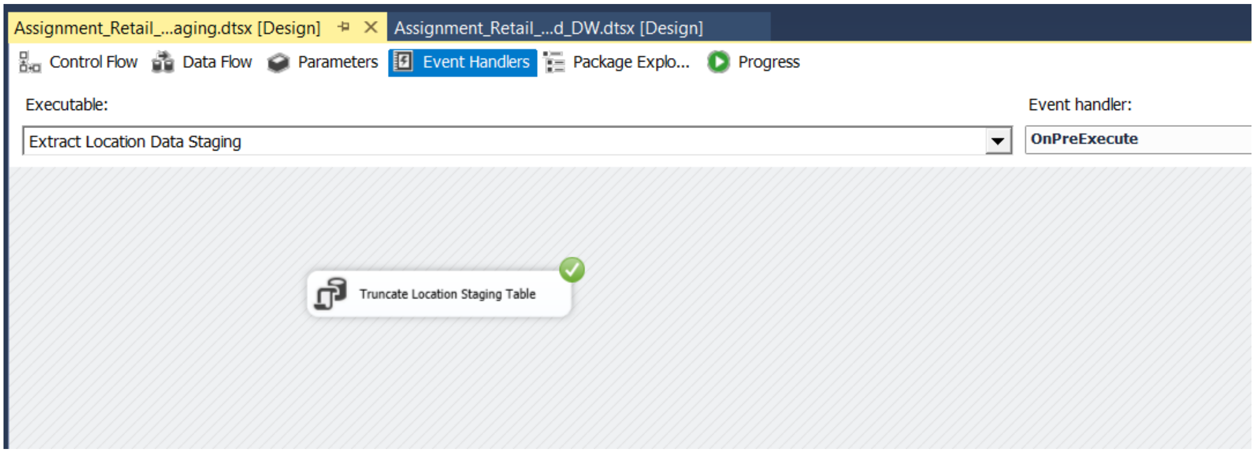
Executable: Extract Flight Data to Staging

Event handler: OnPreExecute


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Truncate Accident Staging Table

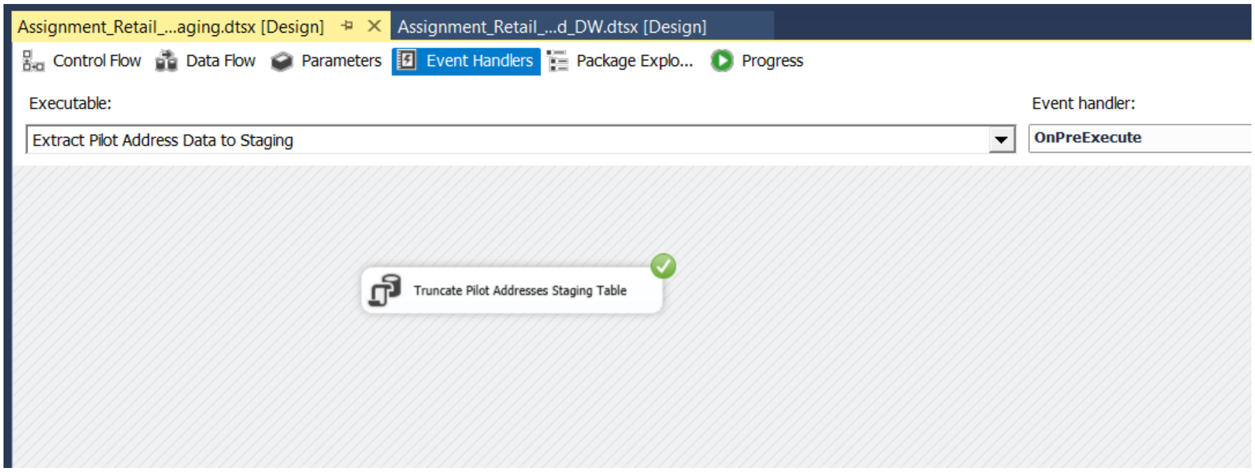
Executable: Extract Accident Data to Staging

Event handler: OnPreExecute


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Truncate Location Staging Table

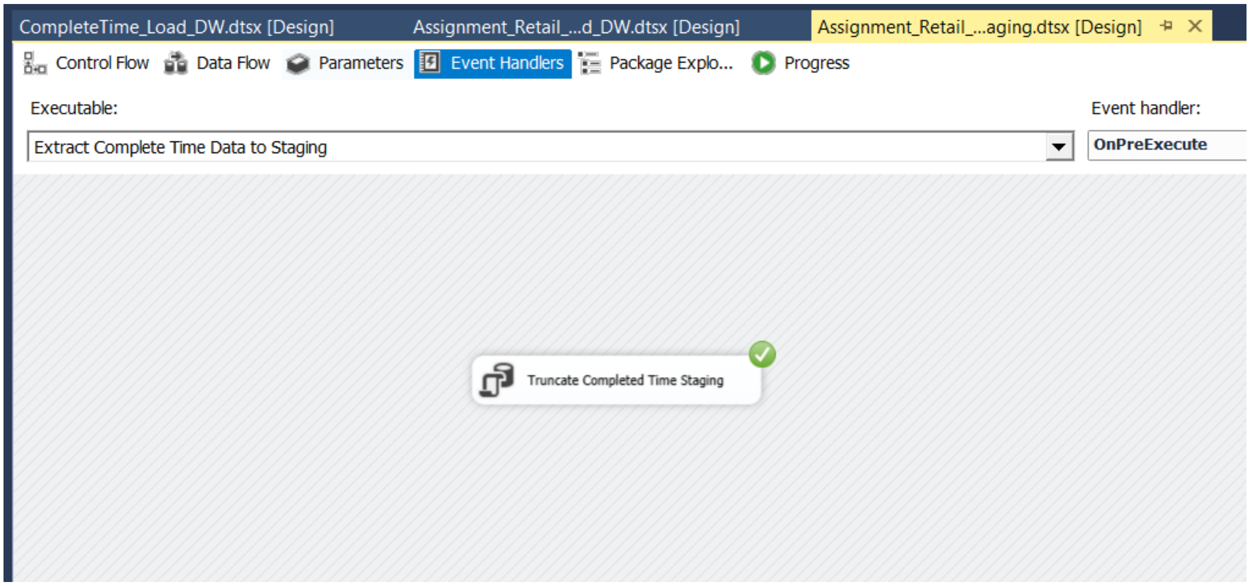
Executable: Extract Location Data Staging

Event handler: OnPreExecute


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Truncate Pilot Addresses Staging Table

Executable: Extract Pilot Address Data to Staging

Event handler: OnPreExecute


CompleteTime_Load_DW.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Assignment_Retail_...aging.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Truncate Completed Time Staging

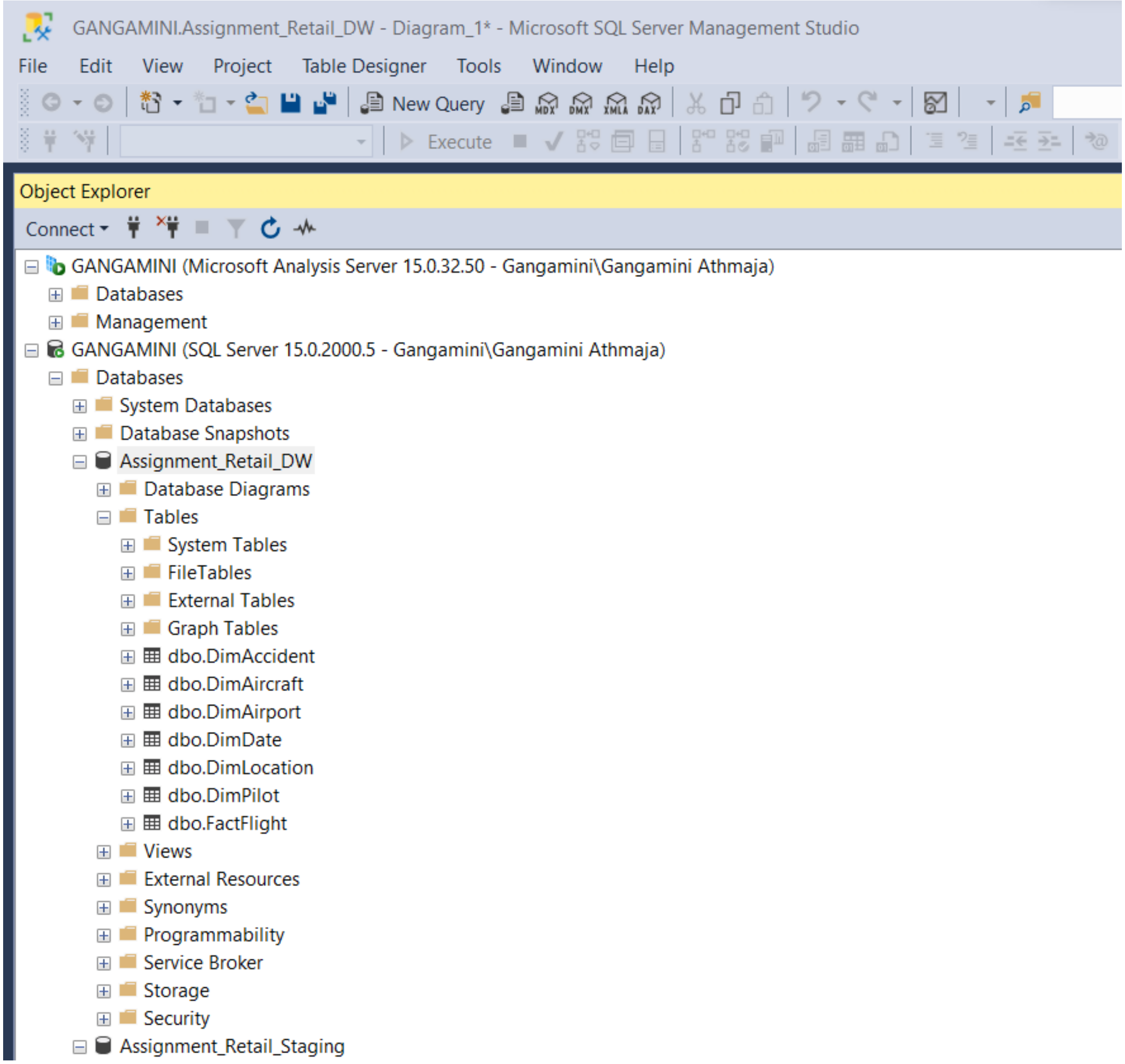
Executable: Extract Complete Time Data to Staging

Event handler: OnPreExecute

3.Transform & Load

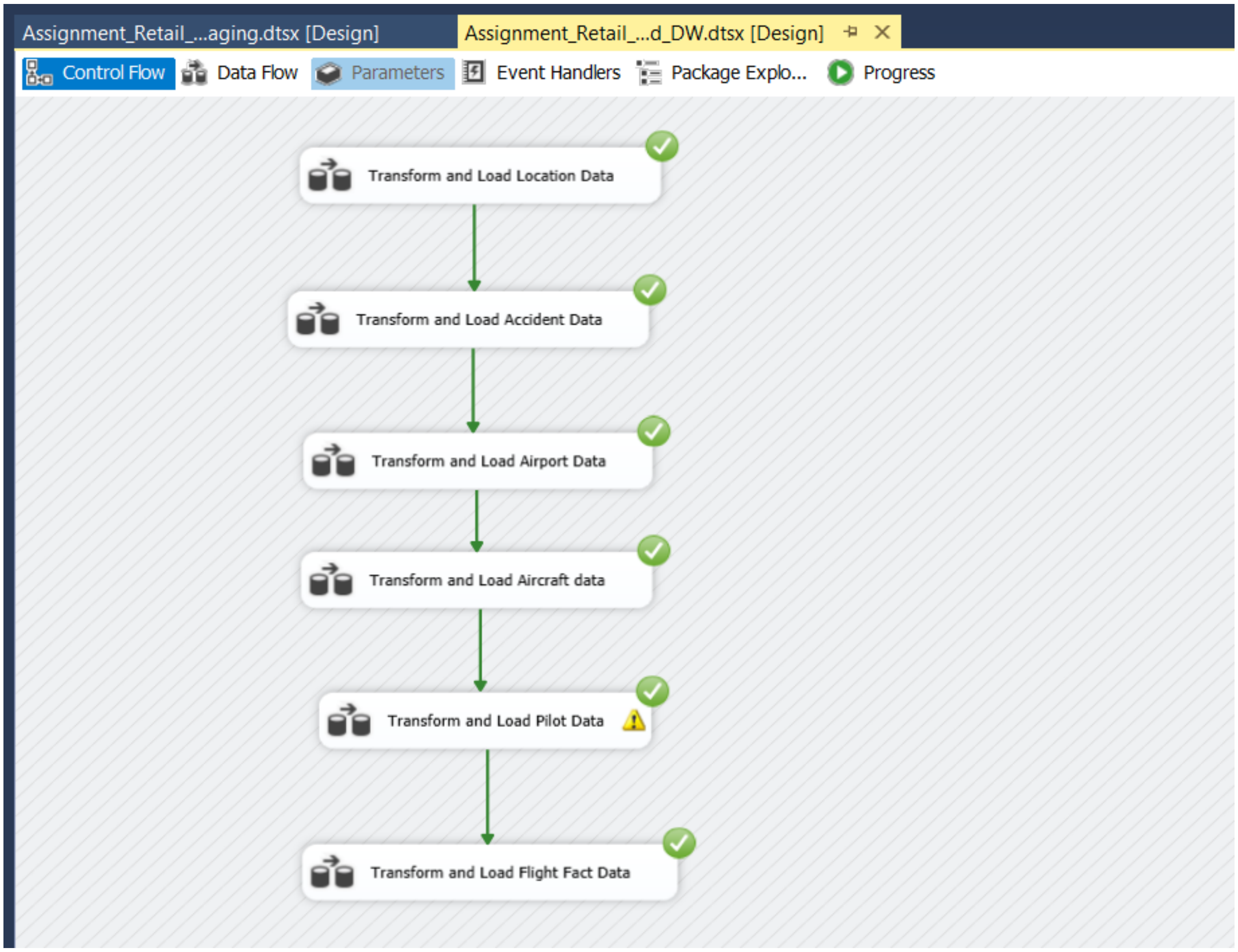
In this step, both the ‘Transform’ and ‘Load’ are done. Firstly, The Dimension tables in the Datawarehouse DB data were created. Then, using the relevant components, data from the staging tables was loaded into the warehouse tables, Assignment_Retail_DW, which contains the below tables,

- **Snapshot of SQL server Data warehouse Database**


GANGAMINI.Assignment_Retail_DW - Diagram_1* - Microsoft SQL Server Management Studio
File Edit View Project Table Designer Tools Window Help
New Query
Execute
Object Explorer
Connect

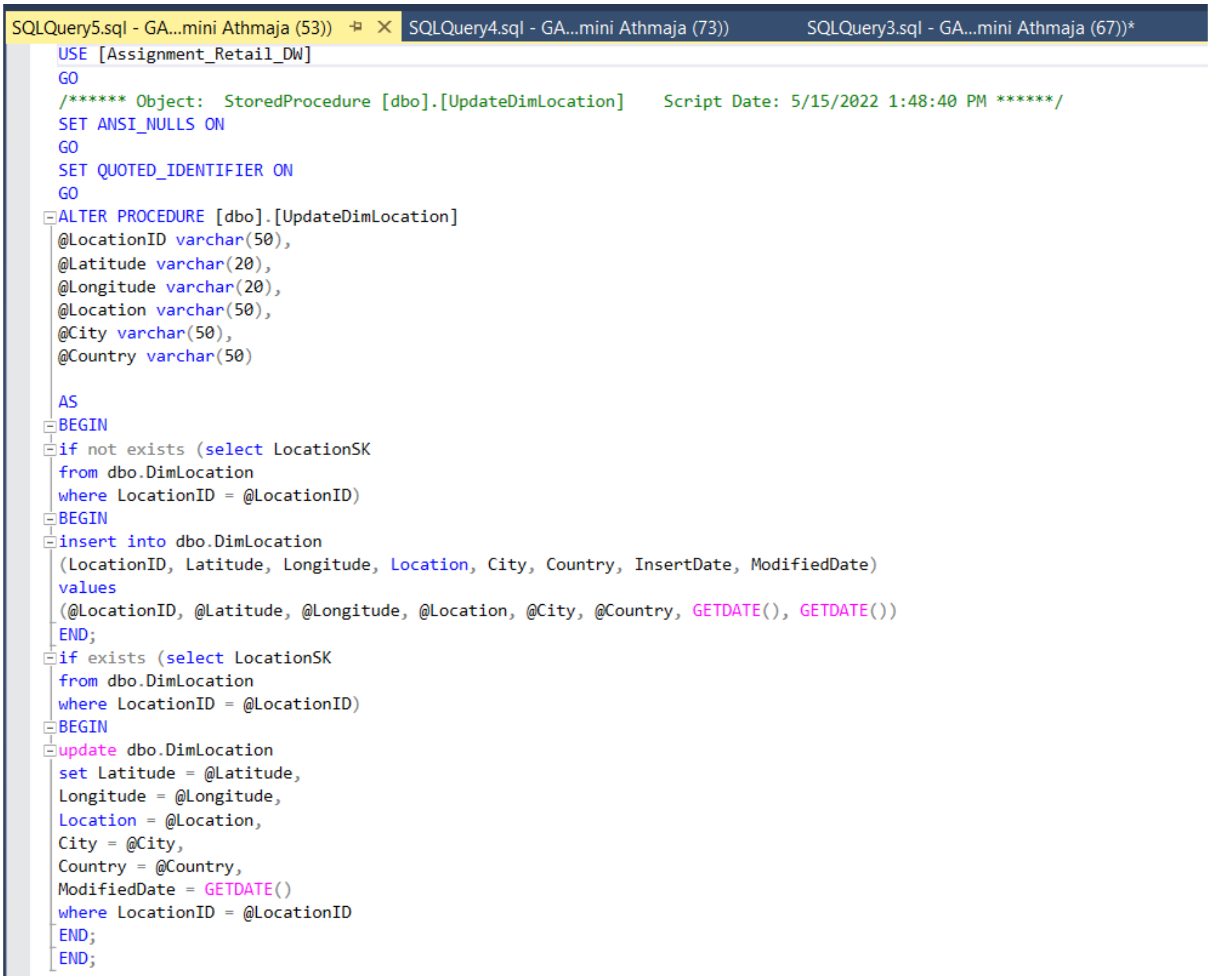
```
GANGAMINI (Microsoft Analysis Server 15.0.32.50 - Gangamini\Gangamini Athmaja)
  Databases
  Management
GANGAMINI (SQL Server 15.0.2000.5 - Gangamini\Gangamini Athmaja)
  Databases
    System Databases
    Database Snapshots
    Assignment_Retail_DW
      Database Diagrams
      Tables
        System Tables
        FileTables
        External Tables
        Graph Tables
        dbo.DimAccident
        dbo.DimAircraft
        dbo.DimAirport
        dbo.DimDate
        dbo.DimLocation
        dbo.DimPilot
        dbo.FactFlight
      Views
      External Resources
      Synonyms
      Programmability
      Service Broker
      Storage
      Security
    Assignment_Retail_Staging
```

- **Snapshot of Visual Studio Control Flow of Extraction**


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Transform and Load Location Data
Transform and Load Accident Data
Transform and Load Airport Data
Transform and Load Aircraft data
Transform and Load Pilot Data
Transform and Load Flight Fact Data

Stored Procedures

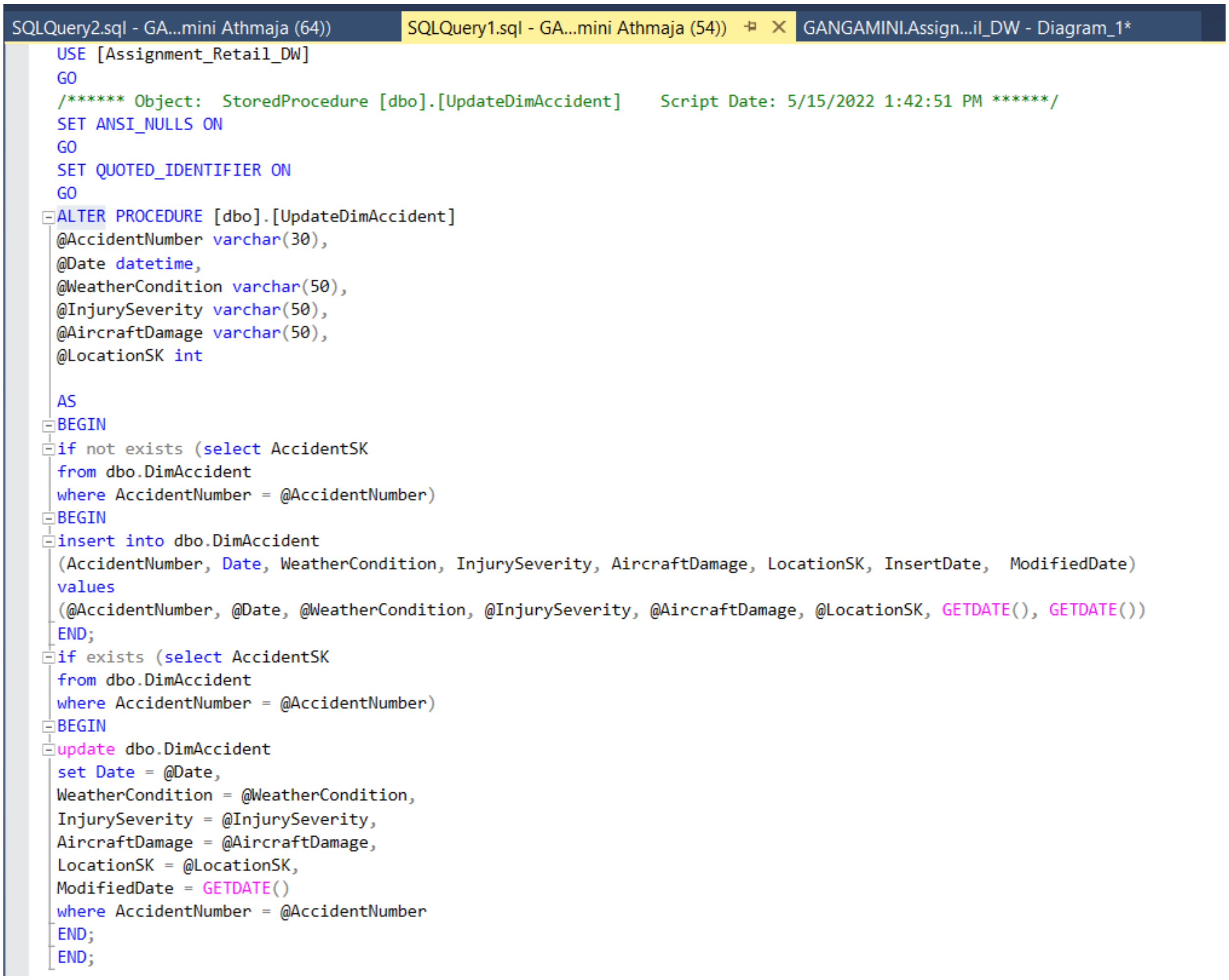
- Location



SQLQuery5.sql - GA...mini Athmaja (53))
SQLQuery4.sql - GA...mini Athmaja (73))
SQLQuery3.sql - GA...mini Athmaja (67))*

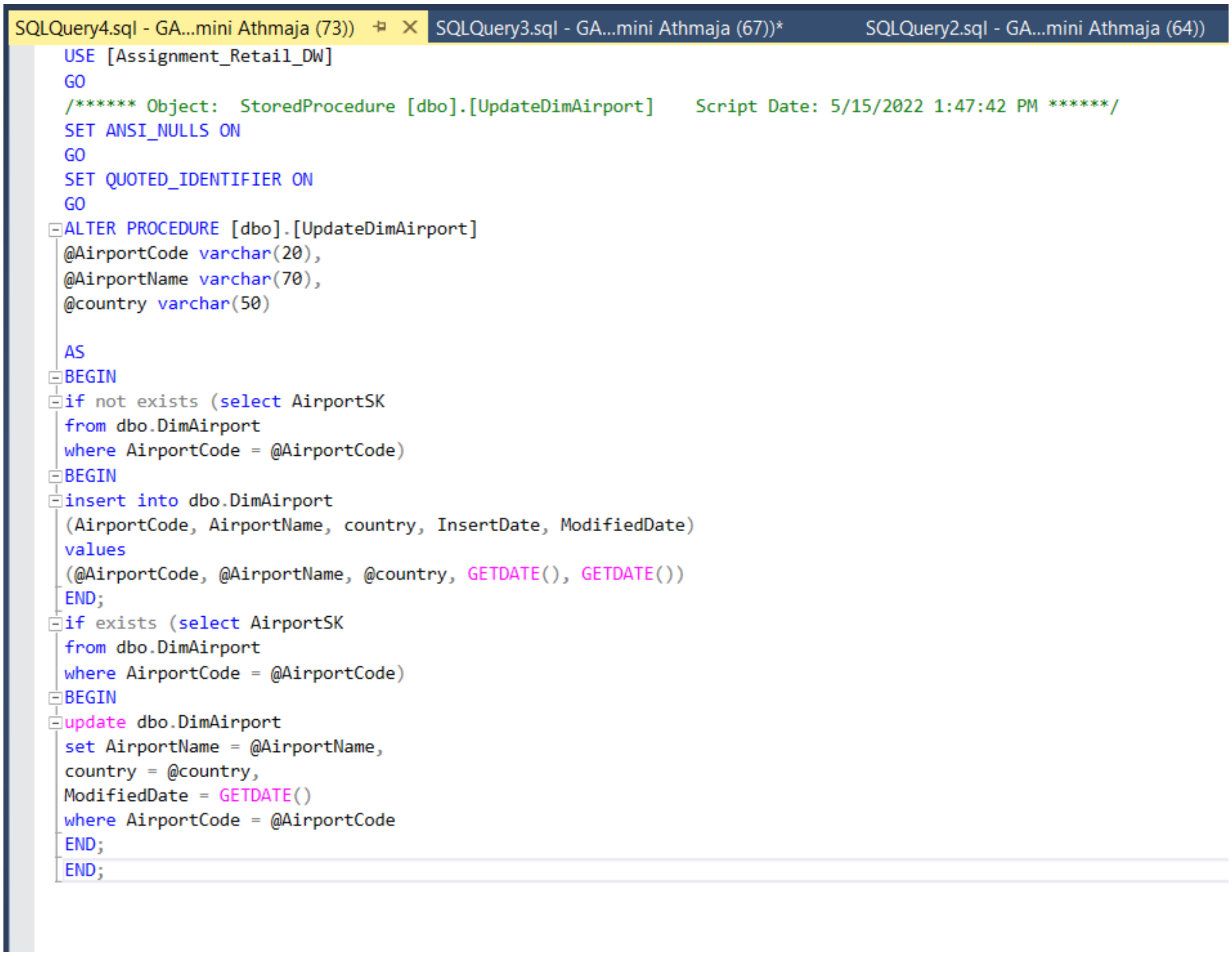
```sql
USE [Assignment_Retail_DW]
GO
/****** Object:  StoredProcedure [dbo].[UpdateDimLocation]    Script Date: 5/15/2022 1:48:40 PM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
ALTER PROCEDURE [dbo].[UpdateDimLocation]
@LocationID varchar(50),
@Latitude varchar(20),
@Longitude varchar(20),
@Location varchar(50),
@City varchar(50),
@Country varchar(50)

AS
BEGIN
if not exists (select LocationSK
from dbo.DimLocation
where LocationID = @LocationID)
BEGIN
insert into dbo.DimLocation
(LocationID, Latitude, Longitude, Location, City, Country, InsertDate, ModifiedDate)
values
(@LocationID, @Latitude, @Longitude, @Location, @City, @Country, GETDATE(), GETDATE())
END;
if exists (select LocationSK
from dbo.DimLocation
where LocationID = @LocationID)
BEGIN
update dbo.DimLocation
set Latitude = @Latitude,
Longitude = @Longitude,
Location = @Location,
City = @City,
Country = @Country,
ModifiedDate = GETDATE()
where LocationID = @LocationID
END;
END;
```

- Accident


SQLQuery2.sql - GA...mini Athmaja (64))
SQLQuery1.sql - GA...mini Athmaja (54))
GANGAMINI.Assign...il_DW - Diagram_1*

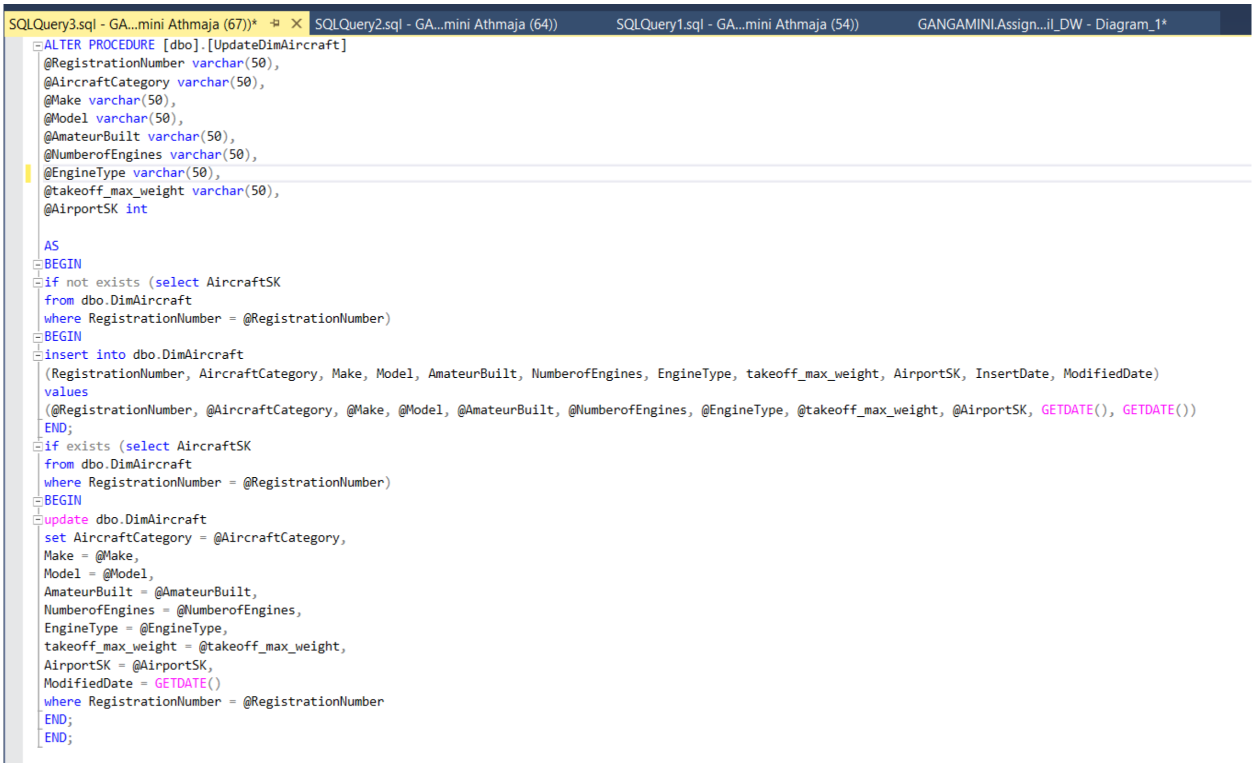
```
USE [Assignment_Retail_DW]
GO
/****** Object:  StoredProcedure [dbo].[UpdateDimAccident]    Script Date: 5/15/2022 1:42:51 PM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
ALTER PROCEDURE [dbo].[UpdateDimAccident]
@AccidentNumber varchar(30),
@Date datetime,
@WeatherCondition varchar(50),
@InjurySeverity varchar(50),
@AircraftDamage varchar(50),
@LocationSK int

AS
BEGIN
if not exists (select AccidentSK
from dbo.DimAccident
where AccidentNumber = @AccidentNumber)
BEGIN
insert into dbo.DimAccident
(AccidentNumber, Date, WeatherCondition, InjurySeverity, AircraftDamage, LocationSK, InsertDate,  ModifiedDate)
values
(@AccidentNumber, @Date, @WeatherCondition, @InjurySeverity, @AircraftDamage, @LocationSK, GETDATE(), GETDATE())
END;
if exists (select AccidentSK
from dbo.DimAccident
where AccidentNumber = @AccidentNumber)
BEGIN
update dbo.DimAccident
set Date = @Date,
WeatherCondition = @WeatherCondition,
InjurySeverity = @InjurySeverity,
AircraftDamage = @AircraftDamage,
LocationSK = @LocationSK,
ModifiedDate = GETDATE()
where AccidentNumber = @AccidentNumber
END;
END;
```

- Airport


SQLQuery4.sql - GA...mini Athmaja (73))
SQLQuery3.sql - GA...mini Athmaja (67))*
SQLQuery2.sql - GA...mini Athmaja (64))

```
USE [Assignment_Retail_DW]
GO
/****** Object:  StoredProcedure [dbo].[UpdateDimAirport]    Script Date: 5/15/2022 1:47:42 PM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
ALTER PROCEDURE [dbo].[UpdateDimAirport]
@AirportCode varchar(20),
@AirportName varchar(70),
@country varchar(50)

AS
BEGIN
if not exists (select AirportSK
from dbo.DimAirport
where AirportCode = @AirportCode)
BEGIN
insert into dbo.DimAirport
(AirportCode, AirportName, country, InsertDate, ModifiedDate)
values
(@AirportCode, @AirportName, @country, GETDATE(), GETDATE())
END;
if exists (select AirportSK
from dbo.DimAirport
where AirportCode = @AirportCode)
BEGIN
update dbo.DimAirport
set AirportName = @AirportName,
country = @country,
ModifiedDate = GETDATE()
where AirportCode = @AirportCode
END;
END;
```

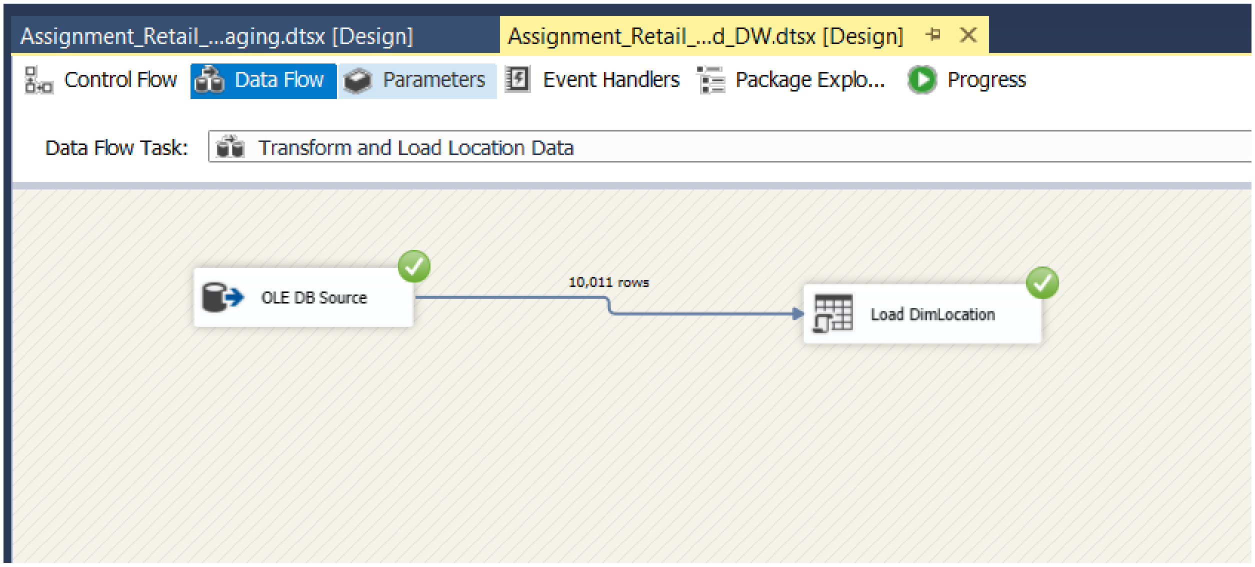
- Aircraft



SQLQuery3.sql - GA...mini Athmaja (67))*
SQLQuery2.sql - GA...mini Athmaja (64))
SQLQuery1.sql - GA...mini Athmaja (54))
GANGAMINI.Assign...il_DW - Diagram_1*

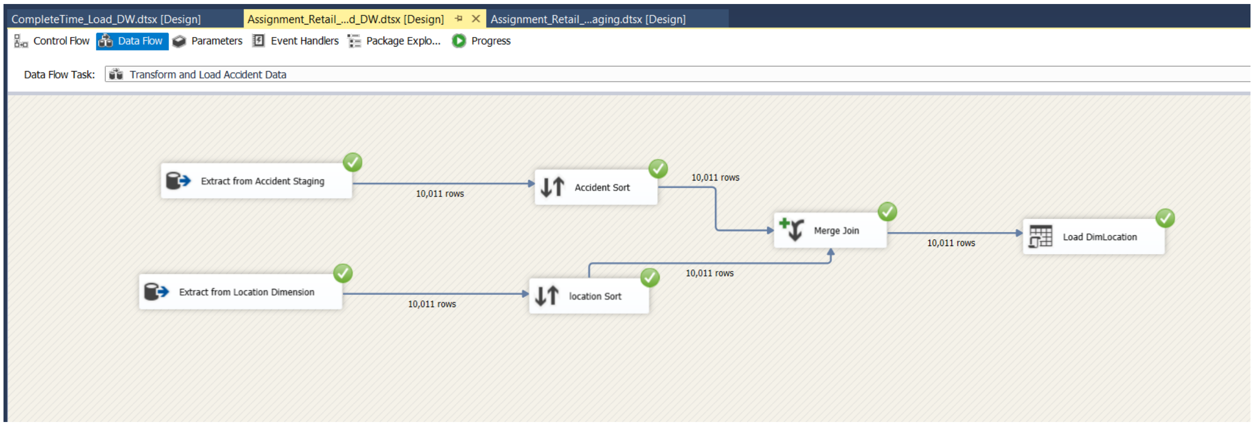
```sql
ALTER PROCEDURE [dbo].[UpdateDimAircraft]
@RegistrationNumber varchar(50),
@AircraftCategory varchar(50),
@Make varchar(50),
@Model varchar(50),
@AmateurBuilt varchar(50),
@NumberofEngines varchar(50),
@EngineType varchar(50),
@takeoff_max_weight varchar(50),
@AirportSK int

AS
BEGIN
if not exists (select AircraftSK
from dbo.DimAircraft
where RegistrationNumber = @RegistrationNumber)
BEGIN
insert into dbo.DimAircraft
(RegistrationNumber, AircraftCategory, Make, Model, AmateurBuilt, NumberofEngines, EngineType, takeoff_max_weight, AirportSK, InsertDate, ModifiedDate)
values
(@RegistrationNumber, @AircraftCategory, @Make, @Model, @AmateurBuilt, @NumberofEngines, @EngineType, @takeoff_max_weight, @AirportSK, GETDATE(), GETDATE())
END;
if exists (select AircraftSK
from dbo.DimAircraft
where RegistrationNumber = @RegistrationNumber)
BEGIN
update dbo.DimAircraft
set AircraftCategory = @AircraftCategory,
Make = @Make,
Model = @Model,
AmateurBuilt = @AmateurBuilt,
NumberofEngines = @NumberofEngines,
EngineType = @EngineType,
takeoff_max_weight = @takeoff_max_weight,
AirportSK = @AirportSK,
ModifiedDate = GETDATE()
where RegistrationNumber = @RegistrationNumber
END;
END;
```

- **Transform and Load Location Data**



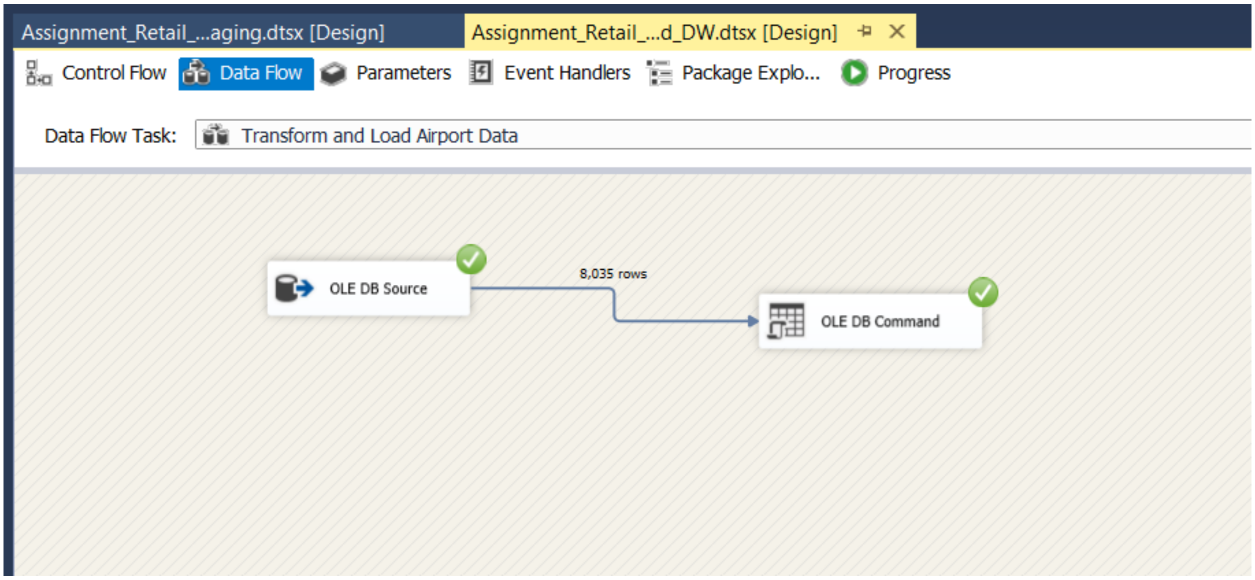
Data Flow Task: Transform and Load Location Data

- **Transform and Load Accident Data**



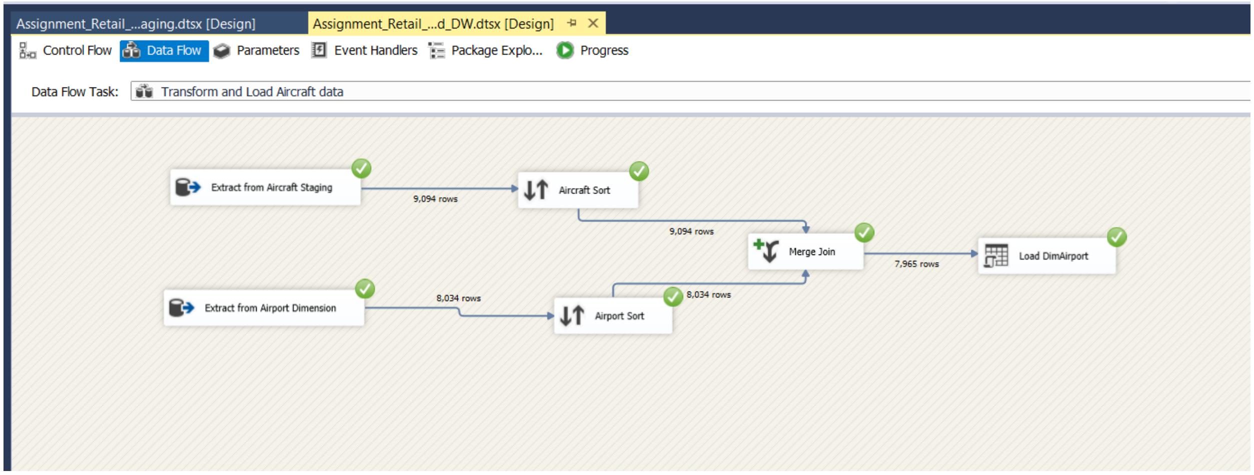
Data Flow Task: Transform and Load Accident Data

- **Transform and Load Airport Data**



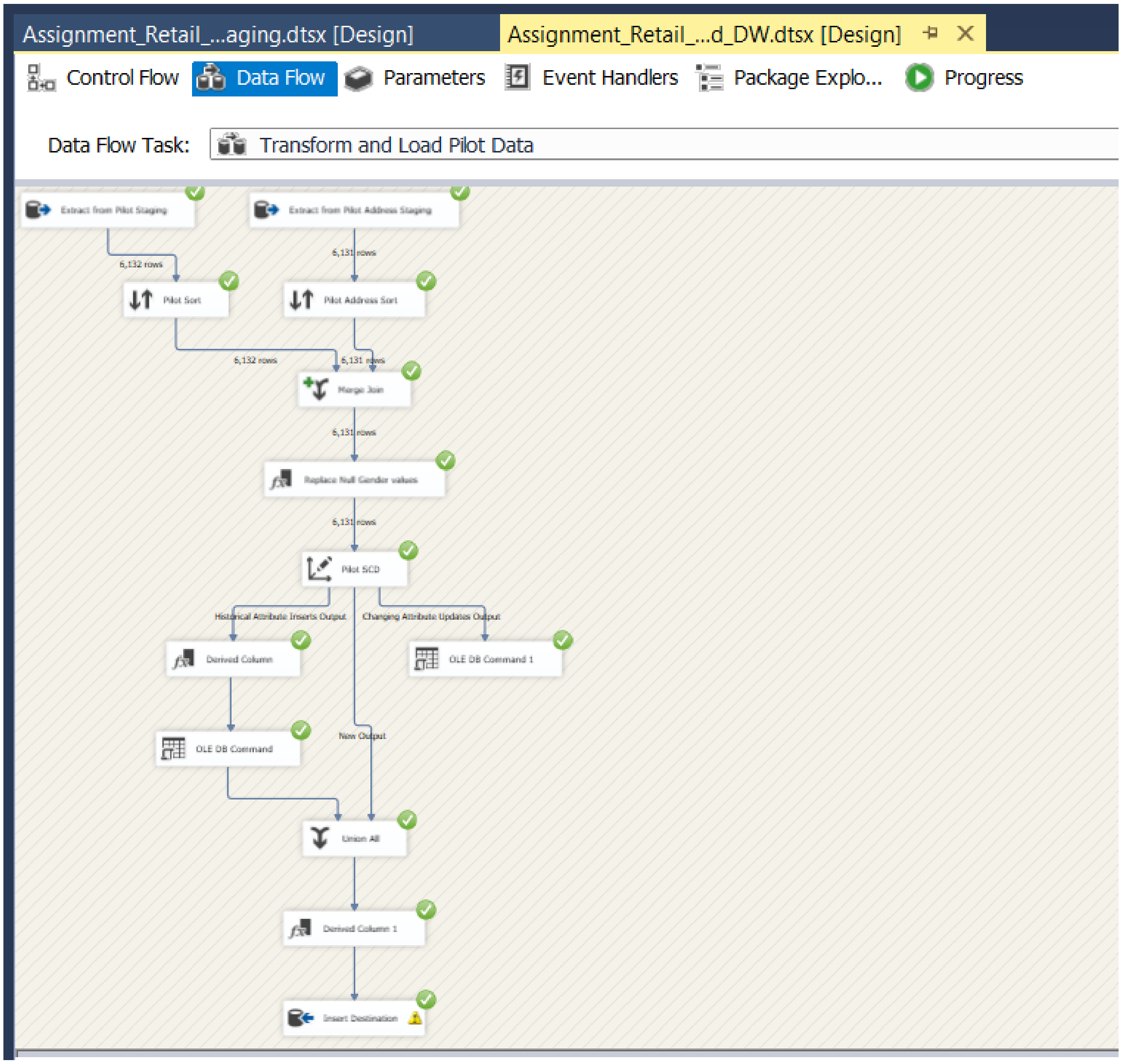
Data Flow Task: Transform and Load Airport Data

- **Transform and Load Aircraft Data**


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress

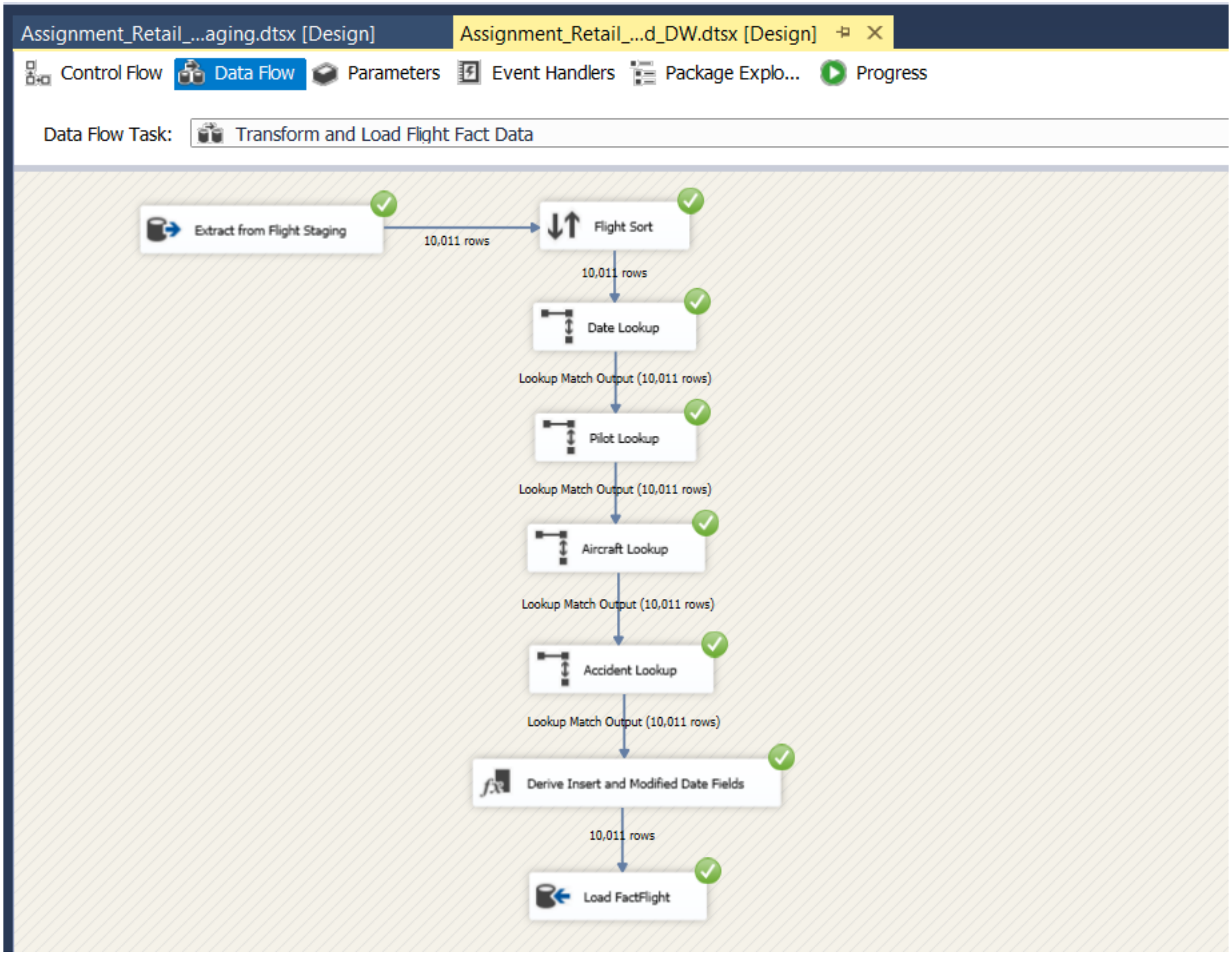
Data Flow Task: Transform and Load Aircraft data

- **Transform and Load Pilot Data**


Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress

Data Flow Task: Transform and Load Pilot Data

- **Transform and Load Flight Fact data**

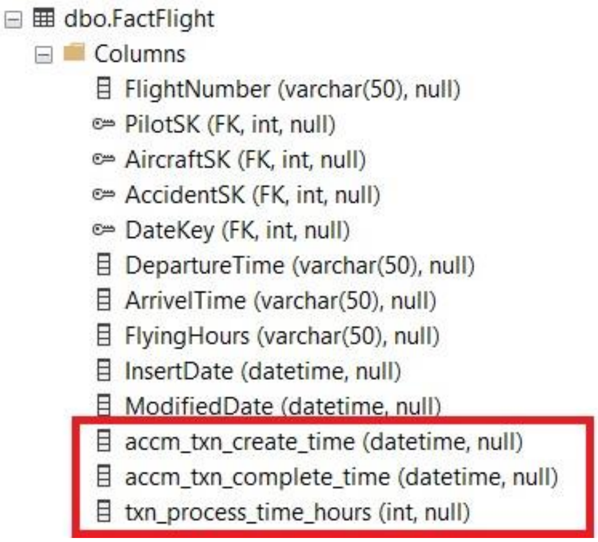


Data Flow Task: Transform and Load Flight Fact Data

Assignment_Retail_...aging.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Extract from Flight Staging
10,011 rows
Flight Sort
10,011 rows
Date Lookup
Lookup Match Output (10,011 rows)
Pilot Lookup
Lookup Match Output (10,011 rows)
Aircraft Lookup
Lookup Match Output (10,011 rows)
Accident Lookup
Lookup Match Output (10,011 rows)
Derive Insert and Modified Date Fields
10,011 rows
Load FactFlight

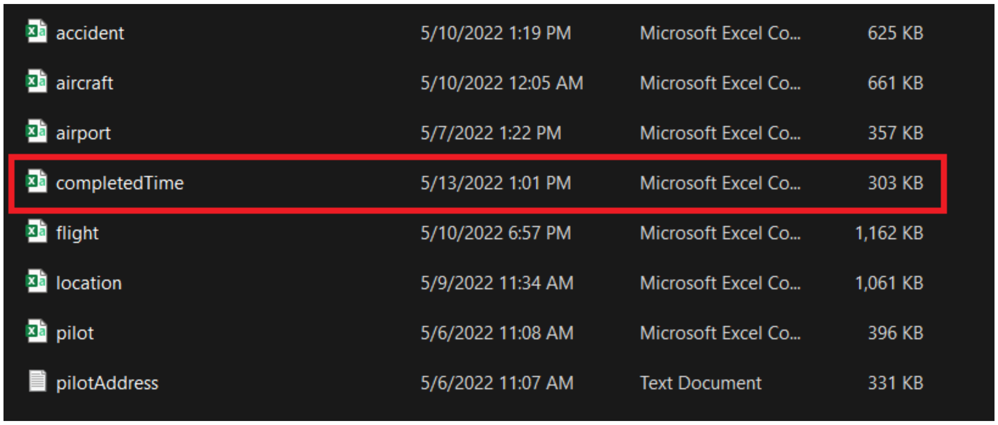
STEP 6: ETL DEVELOPMENT – ACCUMULATING FACT TABLES

- **Extending Fact Table with Additional Columns**



```
dbo.FactFlight
  Columns
    FlightNumber (varchar(50), null)
    PilotSK (FK, int, null)
    AircraftSK (FK, int, null)
    AccidentSK (FK, int, null)
    DateKey (FK, int, null)
    DepartureTime (varchar(50), null)
    ArrivelTime (varchar(50), null)
    FlyingHours (varchar(50), null)
    InsertDate (datetime, null)
    ModifiedDate (datetime, null)
    accm_txn_create_time (datetime, null)
    accm_txn_complete_time (datetime, null)
    txn_process_time_hours (int, null)
```

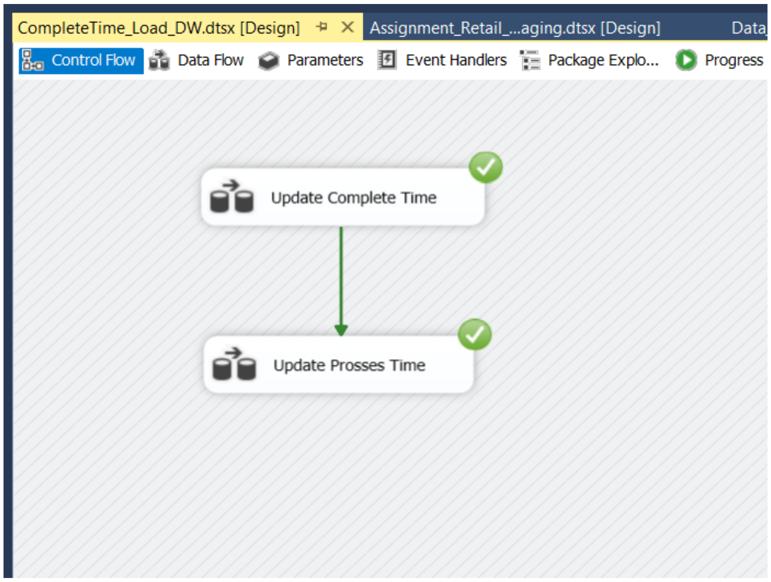
- **Prepare separate data set for complete time**



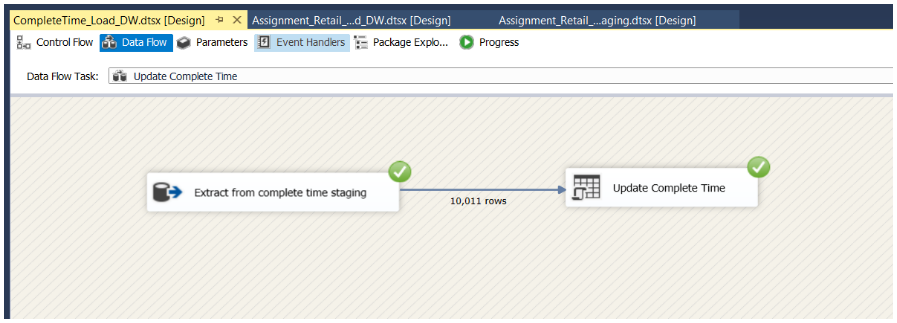
accident	5/10/2022 1:19 PM	Microsoft Excel Co...	625 KB
aircraft	5/10/2022 12:05 AM	Microsoft Excel Co...	661 KB
airport	5/7/2022 1:22 PM	Microsoft Excel Co...	357 KB
completedTime	5/13/2022 1:01 PM	Microsoft Excel Co...	303 KB
flight	5/10/2022 6:57 PM	Microsoft Excel Co...	1,162 KB
location	5/9/2022 11:34 AM	Microsoft Excel Co...	1,061 KB
pilot	5/6/2022 11:08 AM	Microsoft Excel Co...	396 KB
pilotAddress	5/6/2022 11:07 AM	Text Document	331 KB

- **Update Complete Time and Process Time in Fact Table**

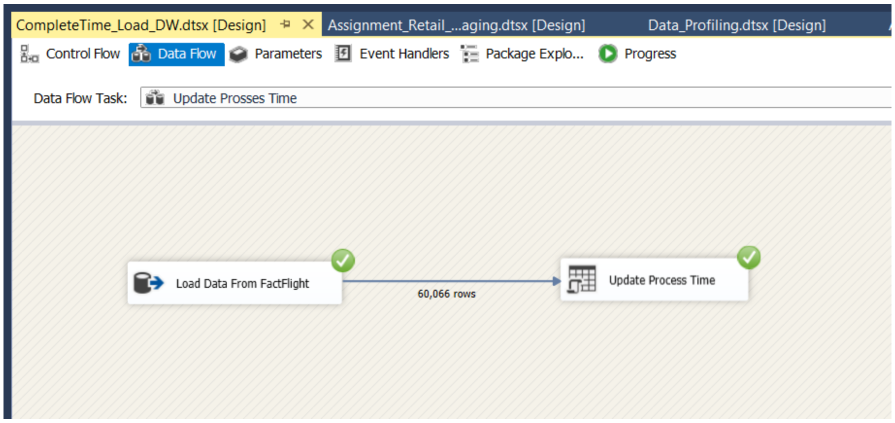
- Control flow



- Data Flows


CompleteTime_Load_DW.dtsx [Design]
Assignment_Retail_...d_DW.dtsx [Design]
Assignment_Retail_...aging.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Extract from complete time staging
10,011 rows
Update Complete Time

Data Flow Task: Update Complete Time


CompleteTime_Load_DW.dtsx [Design]
Assignment_Retail_...aging.dtsx [Design]
Data_Profiling.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Load Data From FactFlight
60,066 rows
Update Process Time

Data Flow Task: Update Prosses Time