Data Warehousing & Business Intelligence

Melbourne Housing Market

Assignment 1

Submitted by:

S.S.D.Peellapitita IT16125094

Submitted to:

MR. Sheron Dinushka

Table of Contents

1. Data set selection

Background

This data set is about Melbourne Housing Market. Melbourne is currently experiencing a housing bubble. May be someone can find a trend or giving a prediction? Which suburbs are the best to buy in. which one are value for money? Where is the expensive side of town?

Content

The data was downloaded for the dates ranging from Jan, 2016 to Nov 27th, 2017

This is the exact link: https://www.kaggle.com/anthonypino/melbourne-housing-market /data

Some Key Details in Dataset

Suburb: Suburb

Address: Address

Rooms: Number of rooms in the house

Price: Price in dollars

Method: S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.

Type: br - bedroom(s); h - house, cottage, villa, semi, terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.

SellerAgent: Real Estate Agent

SoldDate: sold of House

RegionName: General Region (West, North West, North, North east ...etc)

Bedroom : Scraped # of Bedrooms (from different source)

Bathroom: Number of Bathrooms

Car: Number of cars pots

Landsize: Land Size

BuildingArea: Building Size

YearBuilt: Year the house was built

CouncilArea: Governing council for the area

Lattitude: Self explanatory

Longtitude: Self explanitory

2. Preparation of Data Sources

In order to data extraction need to prepare the data sources. From my main data source, I extracted three types of data sources.

1. Database backup (.bak)

2. Text file (.txt)

3. Excel file (.xlsx)

There are five source tables

1. House
2. Location
3. Region
4. Sales
5. Address

Text file: Address

Excel file: House table, Location table, Region table, Sales table

- House :

  HouseID , Rooms, Type, Price, Method, Bedroom, Bathroom, Car, Landsize, BuildingArea, YearBuilt

- Location:

  LocationID, RegionID, Suburb, Postcode, CouncilArea, Latitude, Longitude

- Region:

  RegionID, RegionName

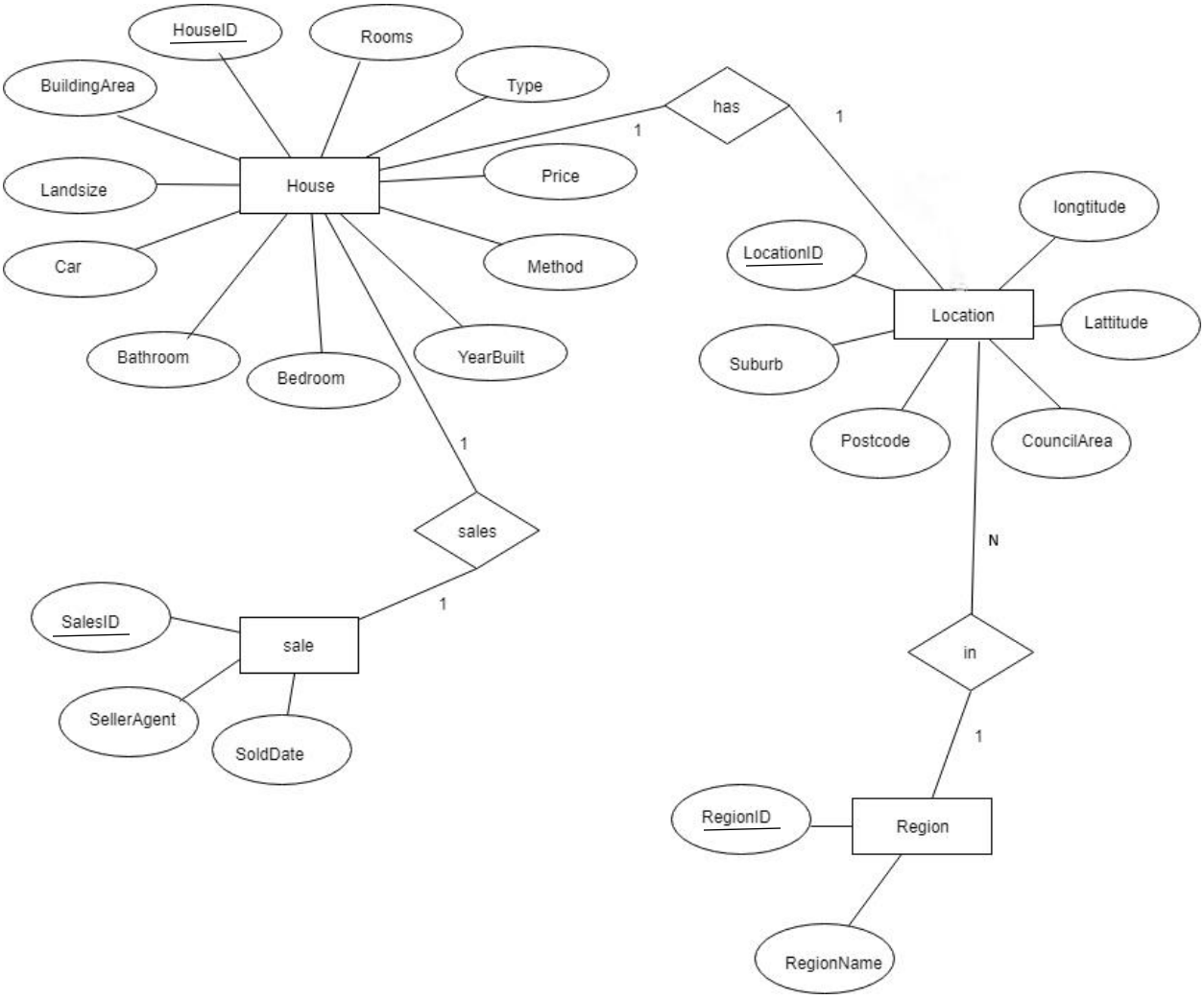
- Sales:

  SalesID, SellerAgent, SoldDate

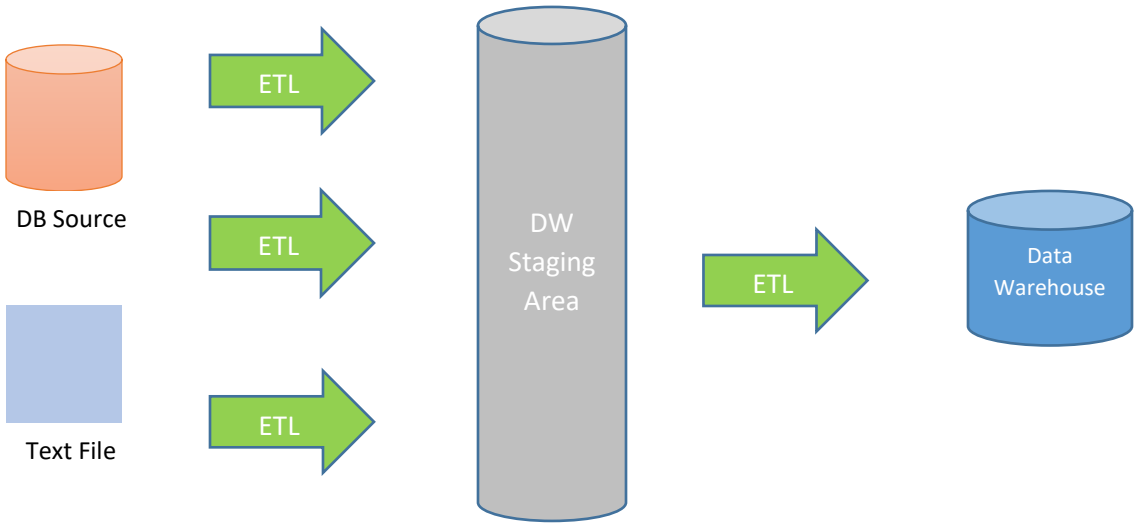
- Address:

  HouseID, Address

ER Diagram


HouseID
Rooms
BuildingArea
Type
has
1
1
House
Landsize
Price
longtitude
LocationID
Car
Method
Location
Lattitude
Suburb
Bathroom
Bedroom
YearBuilt
Postcode
CouncilArea
1
sales
N
SalesID
1
sale
in
SellerAgent
SoldDate
1
RegionID
Region
RegionName

3. Solution Architecture



Address Staging

House Staging

Location Staging

Region Staging

Sales Staging

Architecture Components

- **Data Sources**

Operational System (Transaction)

External sources

- **Extract, Transform, and Load**

Extract – reading data from source systems

Transform – Combine data from multiple sources, De-duplicating

Load – loading data to destination, Surrogate key assignment, Foreign key constraint checks, Indexing

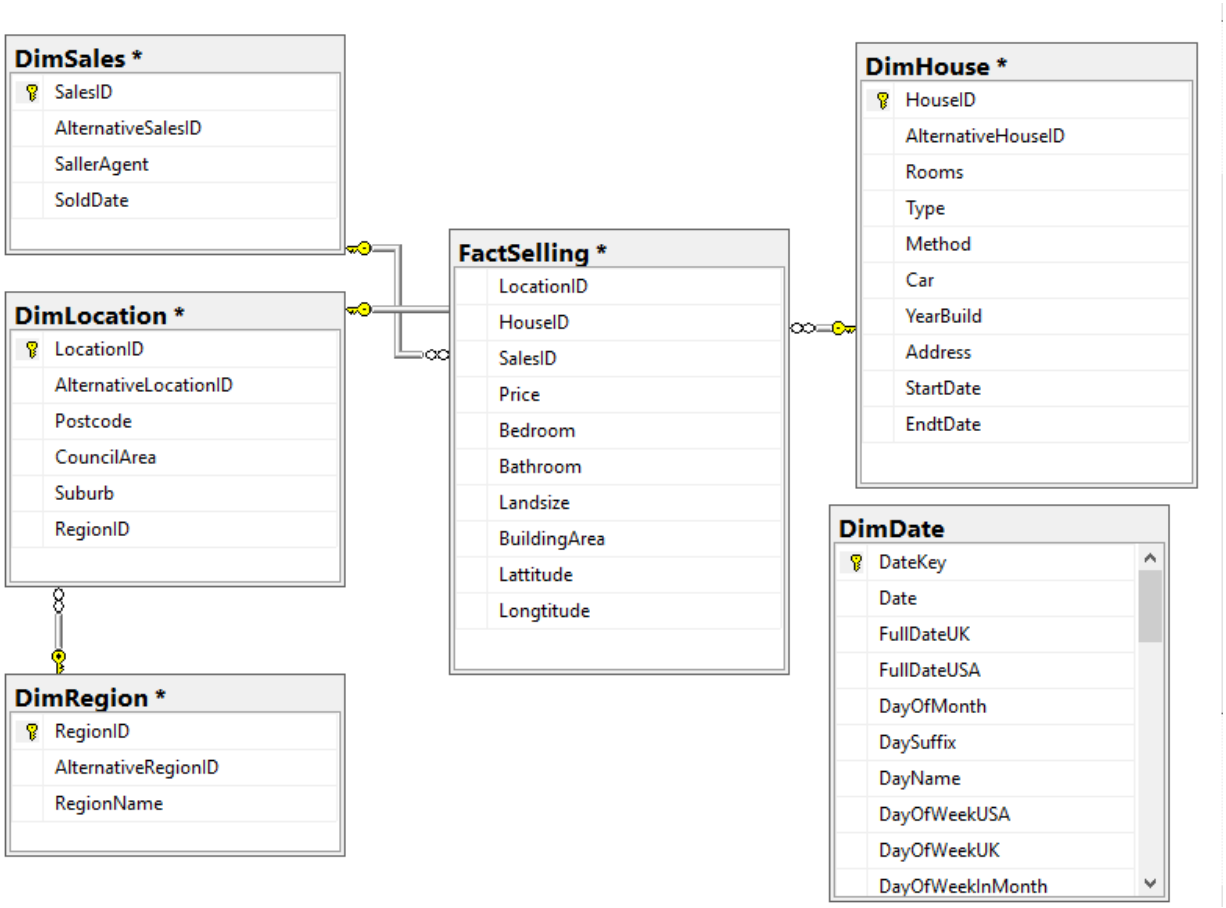
- **Data Warehouse**

EDW vs Data Mart

Dimensional Modeling - Facts & Dimension

Many Schemas -In here used Star schema

4. Data warehouse design & development



5. ETL development

ETL (Extract-Transform-Load)

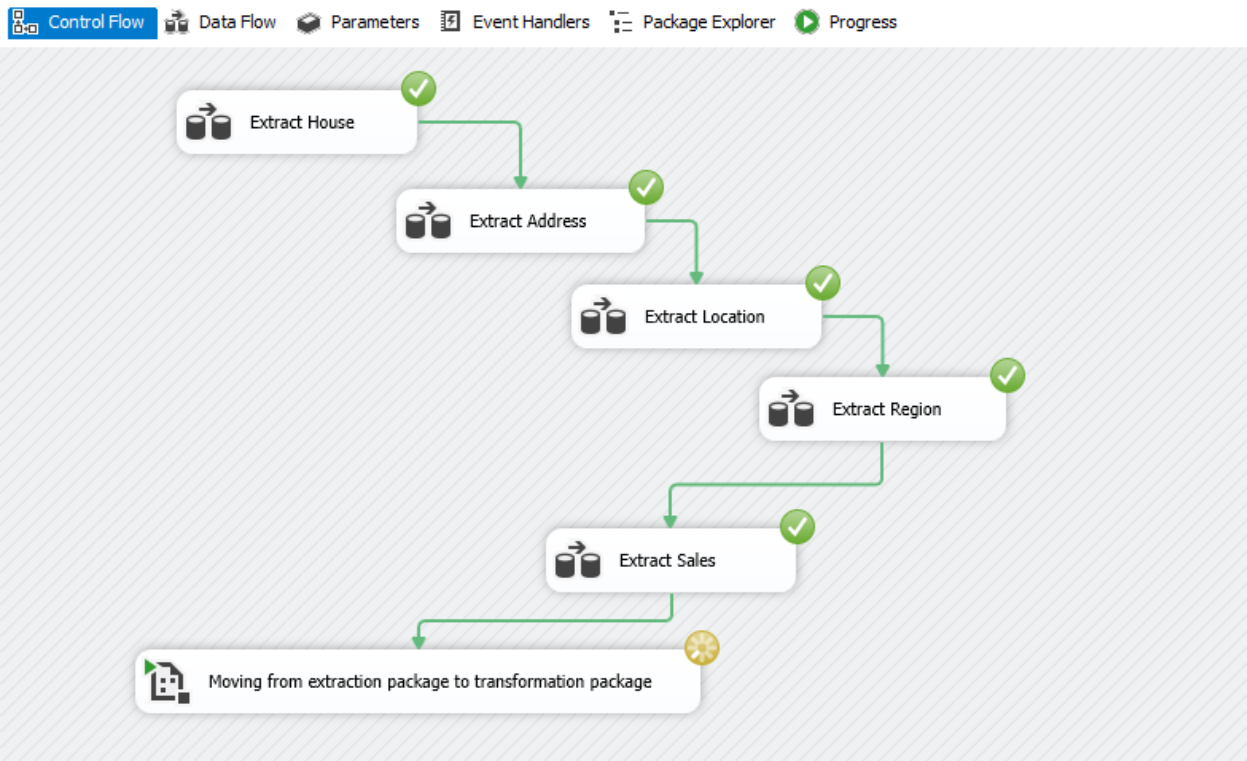
ETL standard for Extract-Transform-Load. ETL covers a process of how the data are loaded from the source system to the data warehouse.

First, the extract function reads data from a specified source database and extracts a desired subset of data. Next, the transform function works with the acquired data - using rules or lookup tables, or creating combinations with other data - to convert it to the desired state. Finally, the load function is used to write the resulting data to a target database.

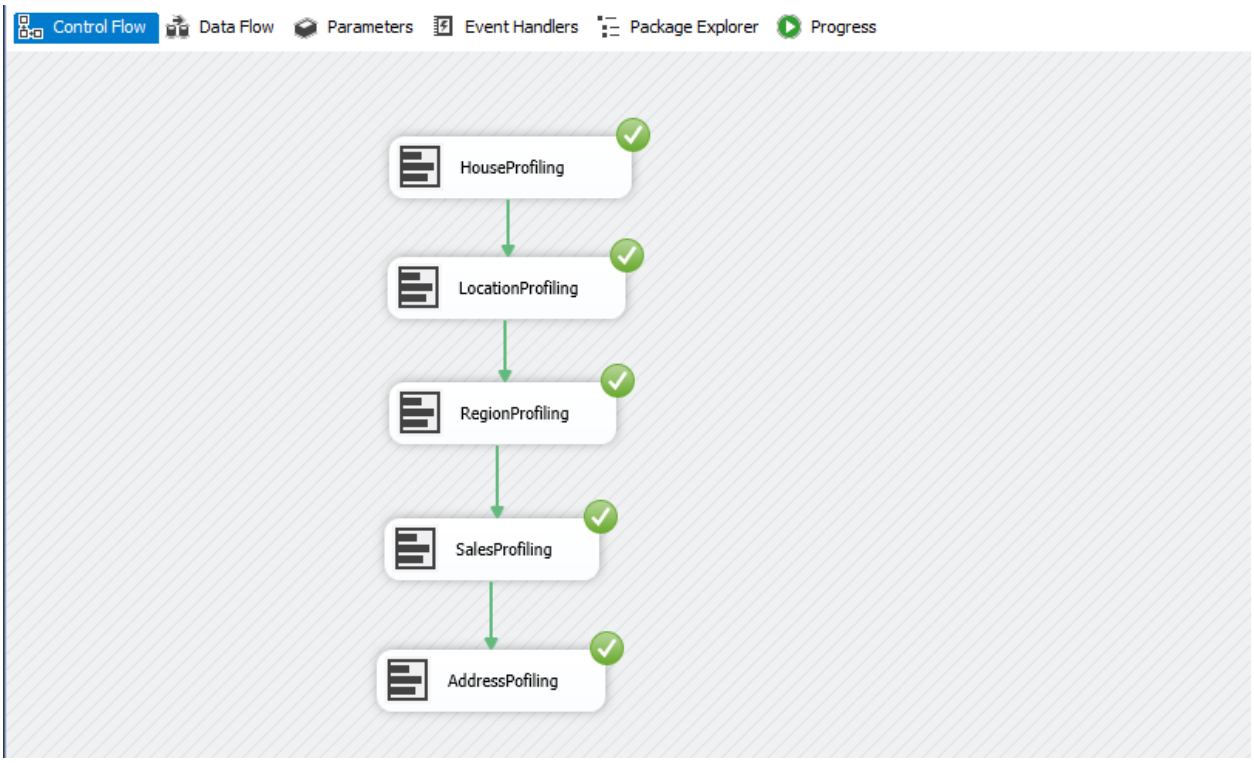
- **Extract**

The main objective of the extract step is to retrieve all the required data from the source system with as little resources as possible.

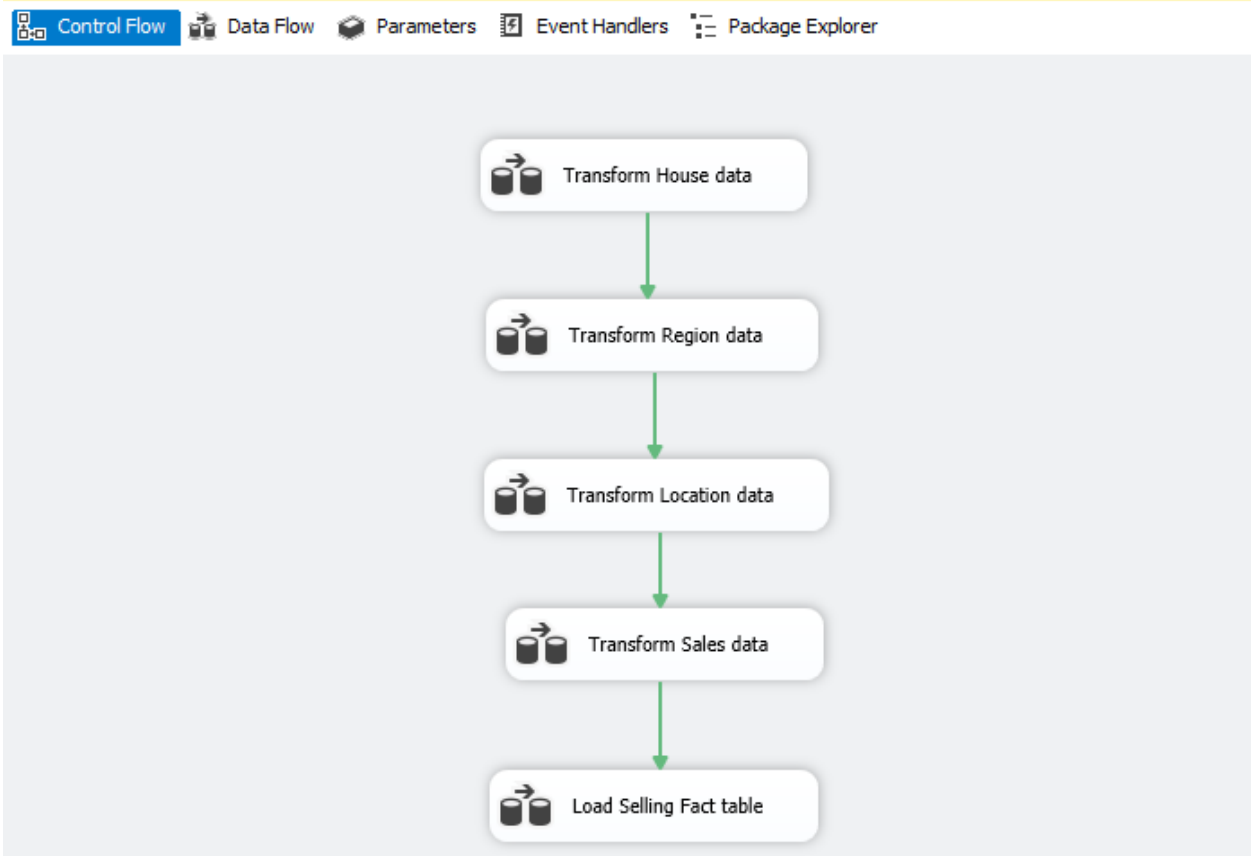
Data Extraction from Sources to Staging Tables



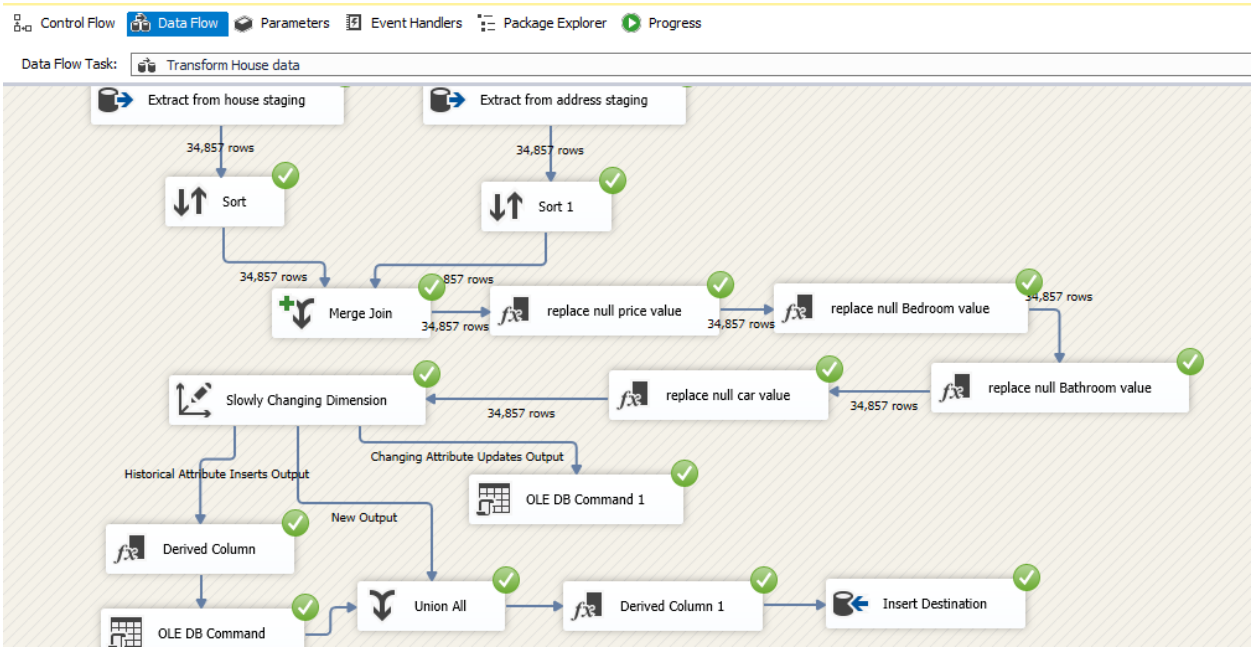
Data Profiling



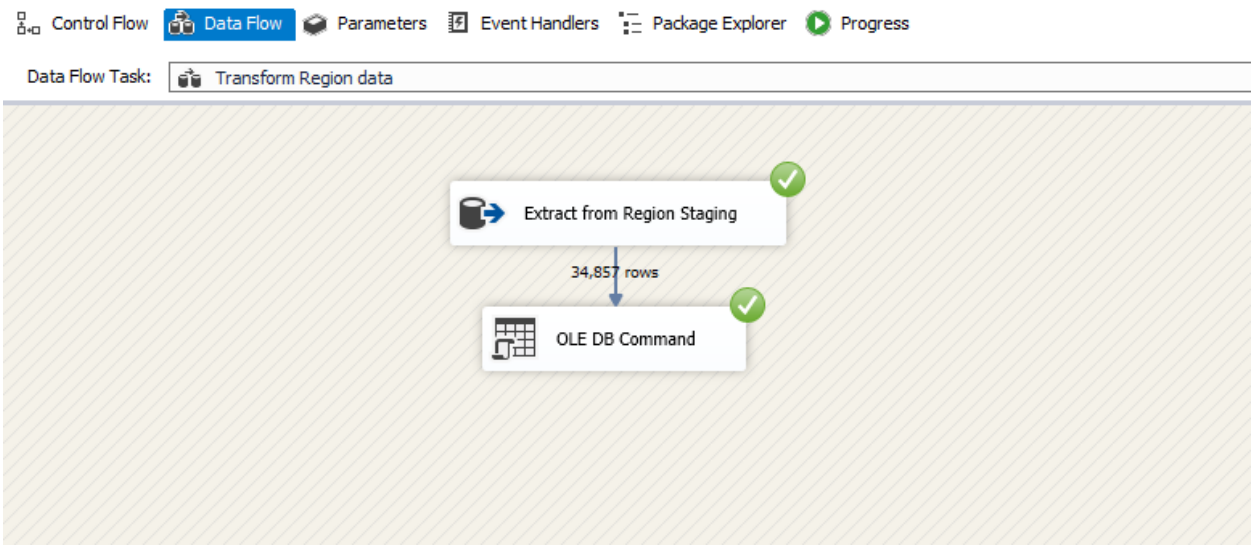
Data Transformation Control Flow



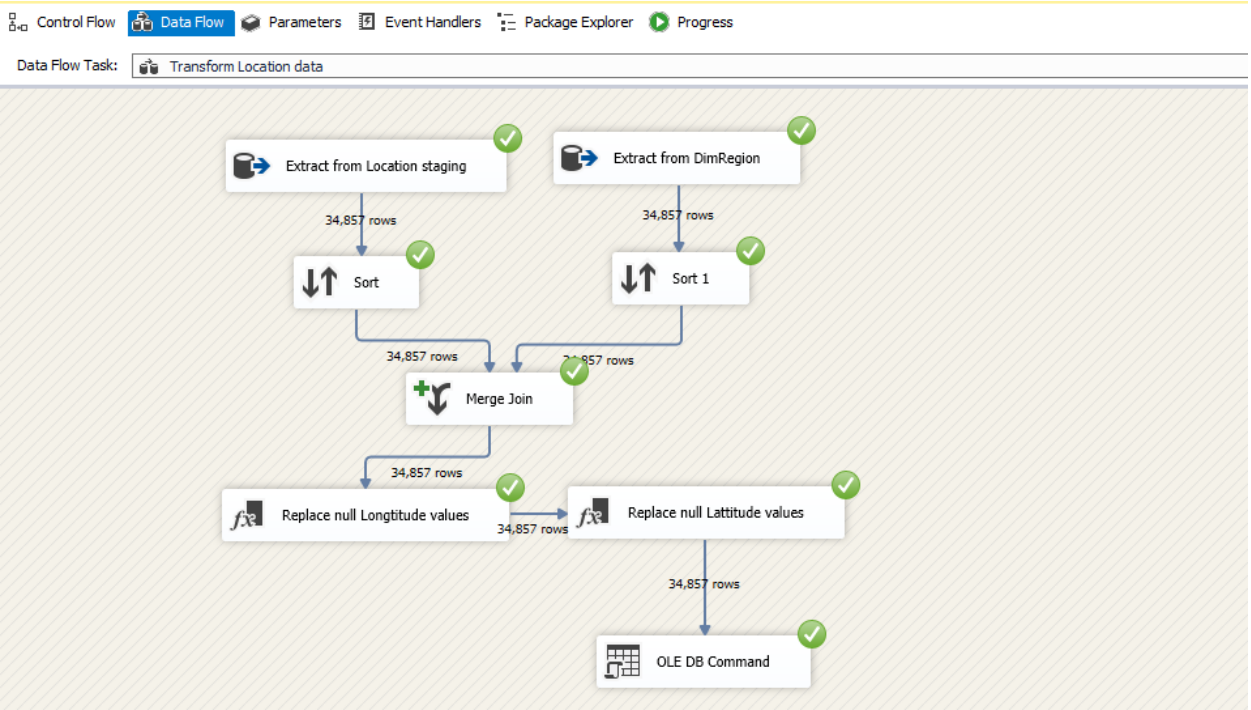
Data Transformation Data Flow of House Staging & Address Staging



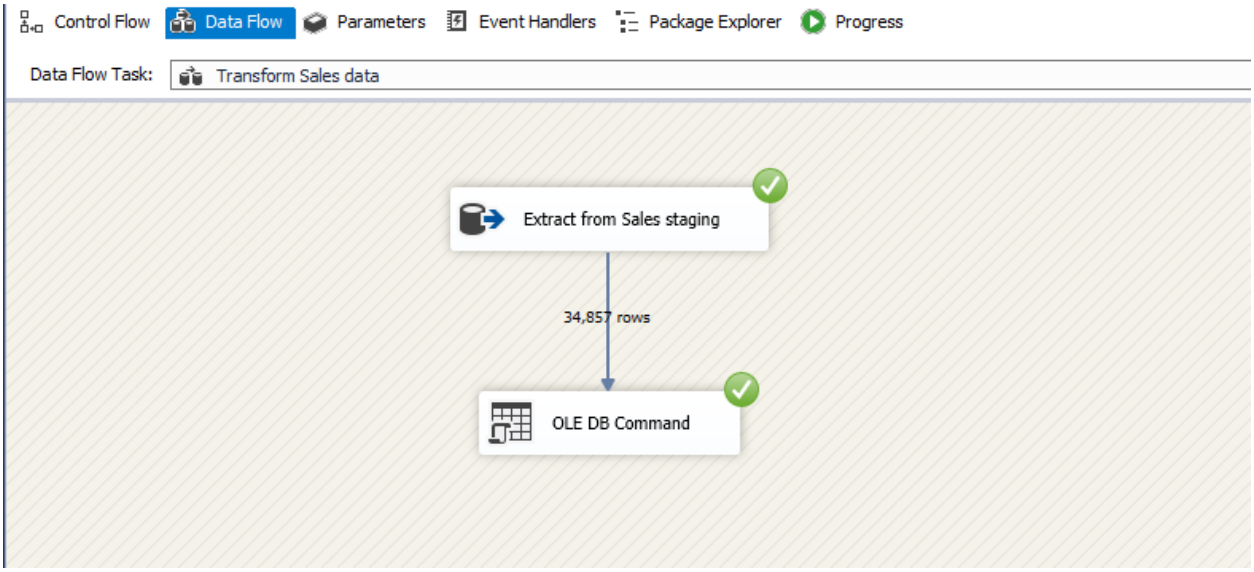
Data Transformation Data Flow of Region Staging



Data Transformation Data Flow of Location Staging & Region staging (Hierarchy)



Data Transformation Data Flow of Sales Staging



Data Flow of Loading Fact Table