

Sri Lanka Institute of Information Technology

**Assignment 1 - Report**

Data Warehousing & Business Intelligence (IT 3021)

2021

**Submitted by:** Jaanvi.S.C.H (IT19801100)

**Submitted on:**  14/05/2021

**Content**--------------------------------------------------------------------------

Data Set Selection………………………………………………………………………………………………………………………..

ER Diagram………………………………………………………………………………………………………………………………….

Preparation of Data Sources………………………………………………………………………………………………………..

Solution Architecture…………………………………………………………………………………………………………………..

Data warehouse design & development……………………………………………………………………………………..

Data Warehouse Data types………………………………………………………………………………………………………..

ETL Development…………………………………………………………………………………………………………………………

* Slowly Changing Dimensin
* Use of Derived Attributes
* Merge , Sort, Union All

Data Profiling……………………………………………………………………………………………………………………………..

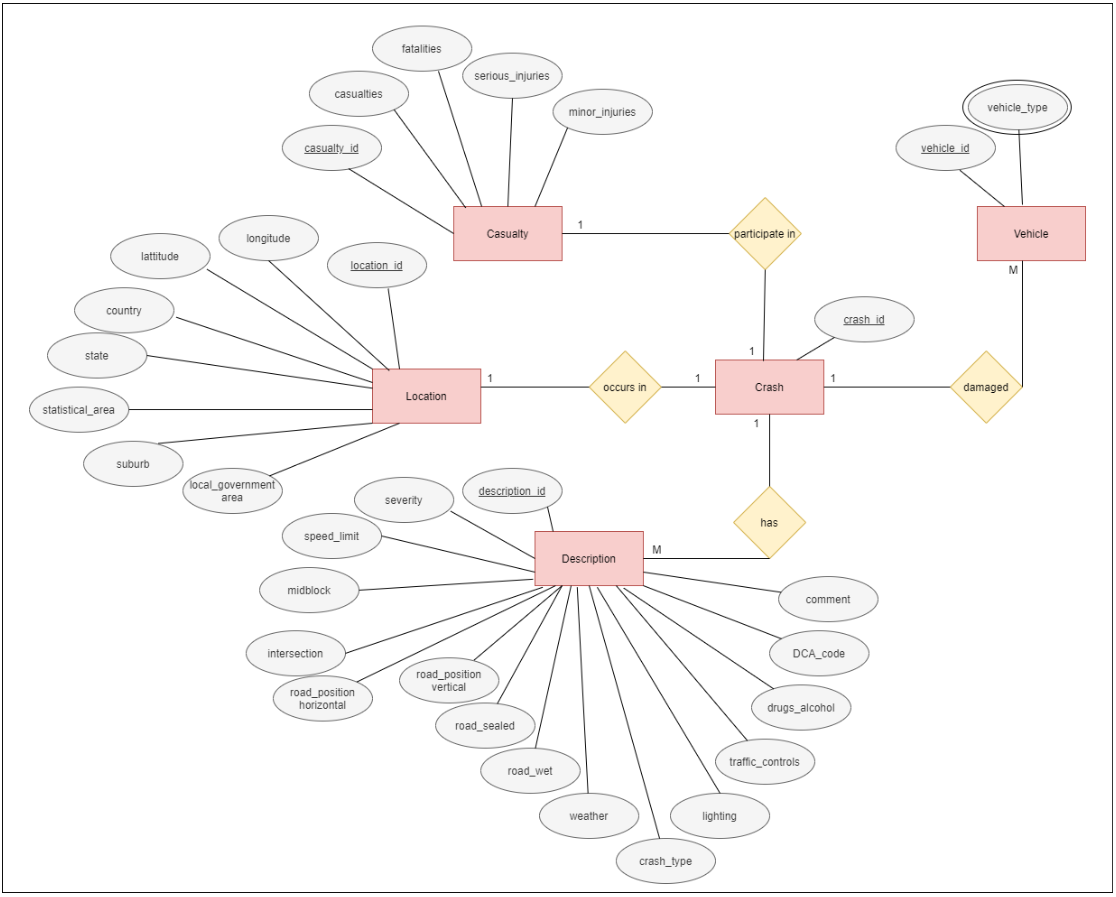
**(Please note : The pdf version of my report loses some images during conversion from word to pdf due to old version mismatches in my laptop so please be kind enough to refer my word document for missing images)**

**Data set selection**

Data Set Link : <https://www.kaggle.com/mgray39/australia-new-zealand-road-crash-dataset>

Australia & New Zealand Road Crash Dataset is a dataset based on where , on what conditions accidents occur and how many casualties were victims in the accidents. This dataset contains 6 CSV tables and more than 1 million data. The data set had sufficient data, according to the needs of the assignment.

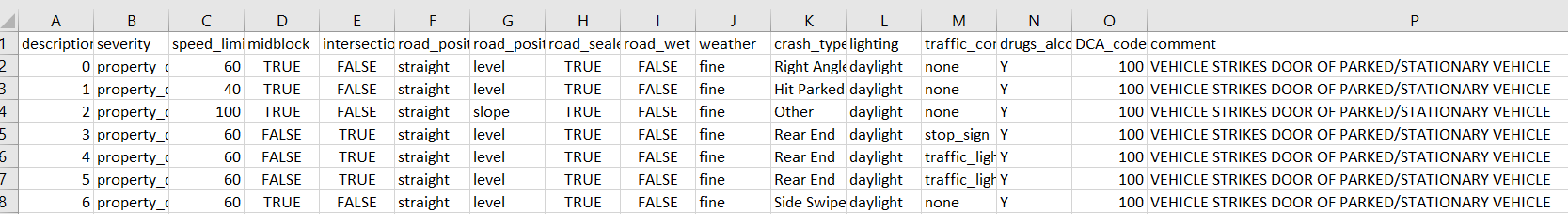
**ER Diagram**



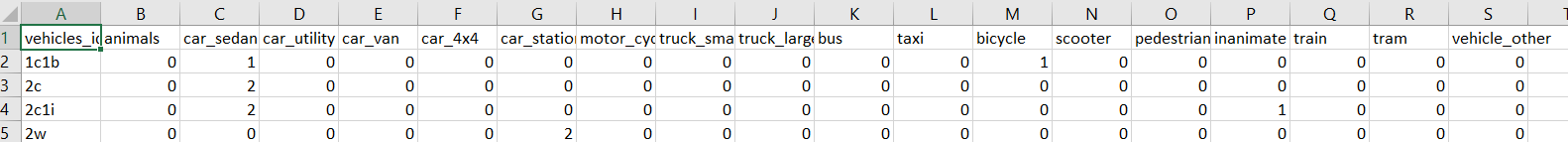
**Preparation of Data Sources**

First of all I converted casualties.csv file into a text file in order to extract data from multiple data sources as per the assignment criteria. Moreover from the dataset I removed data and organized my dataset to fit to 5 years of data.

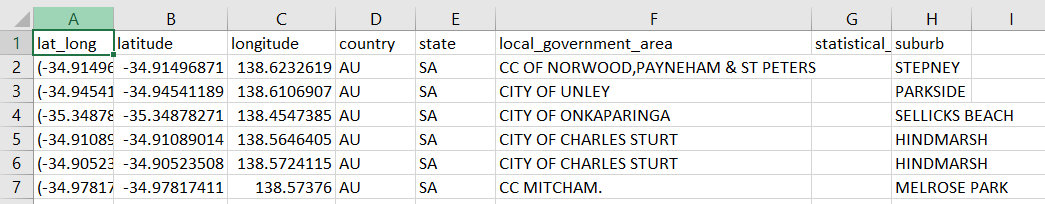
Description.csv



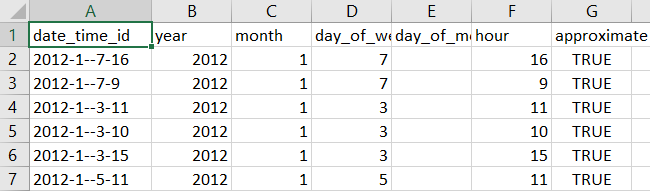
Vehicle.csv



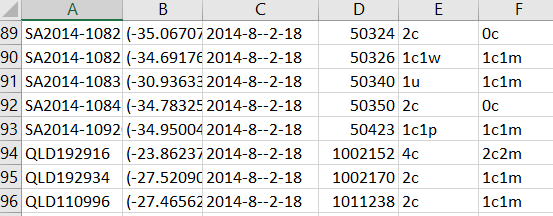
Location.csv

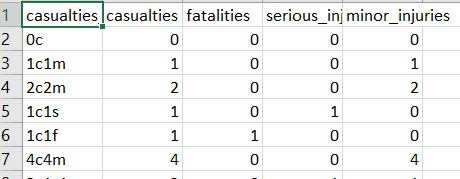


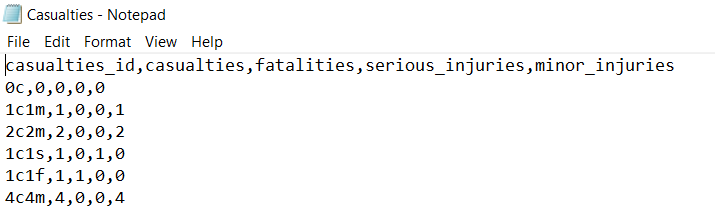
DateTime.csv



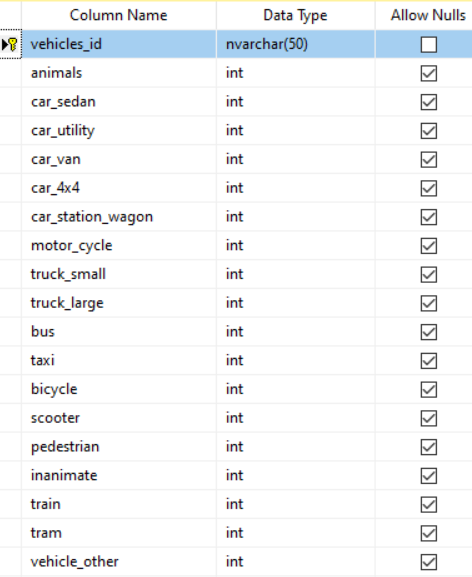
Crash.csv

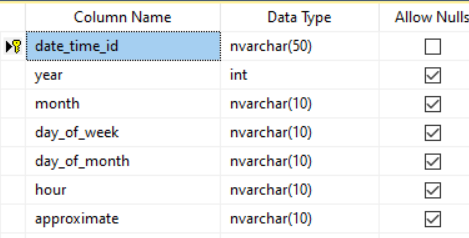


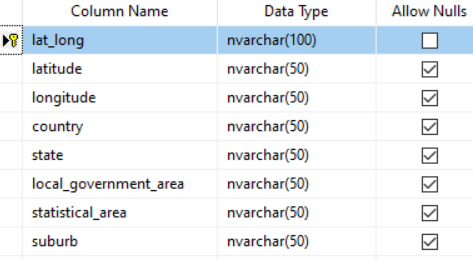
Casualties.csv

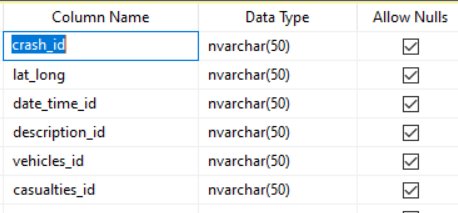
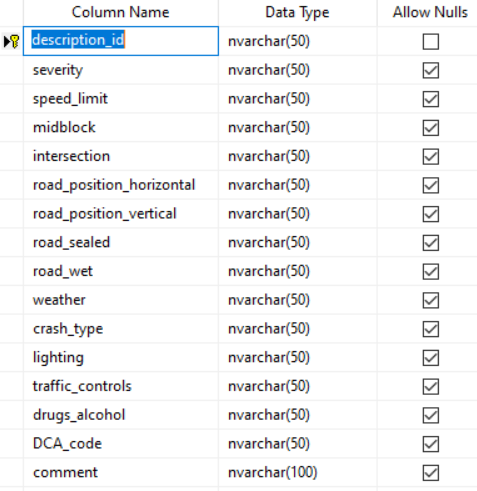
Casualties.txt

**Data types of source tables**

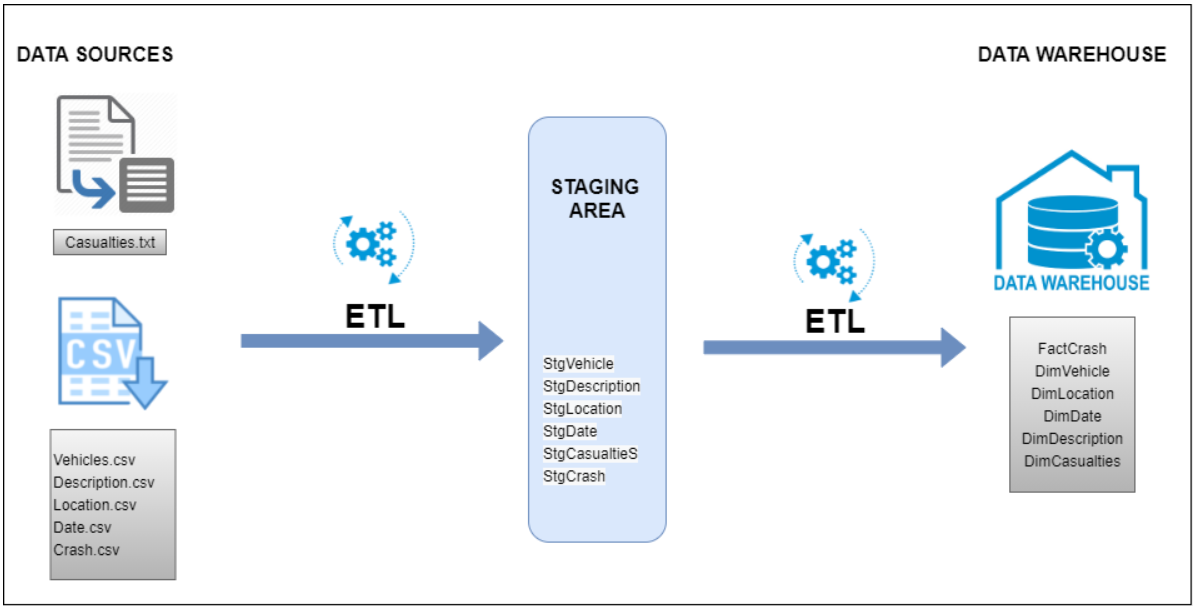
I have created Road\_Crash\_SourceDB Database ad imported the CSV files of source data.

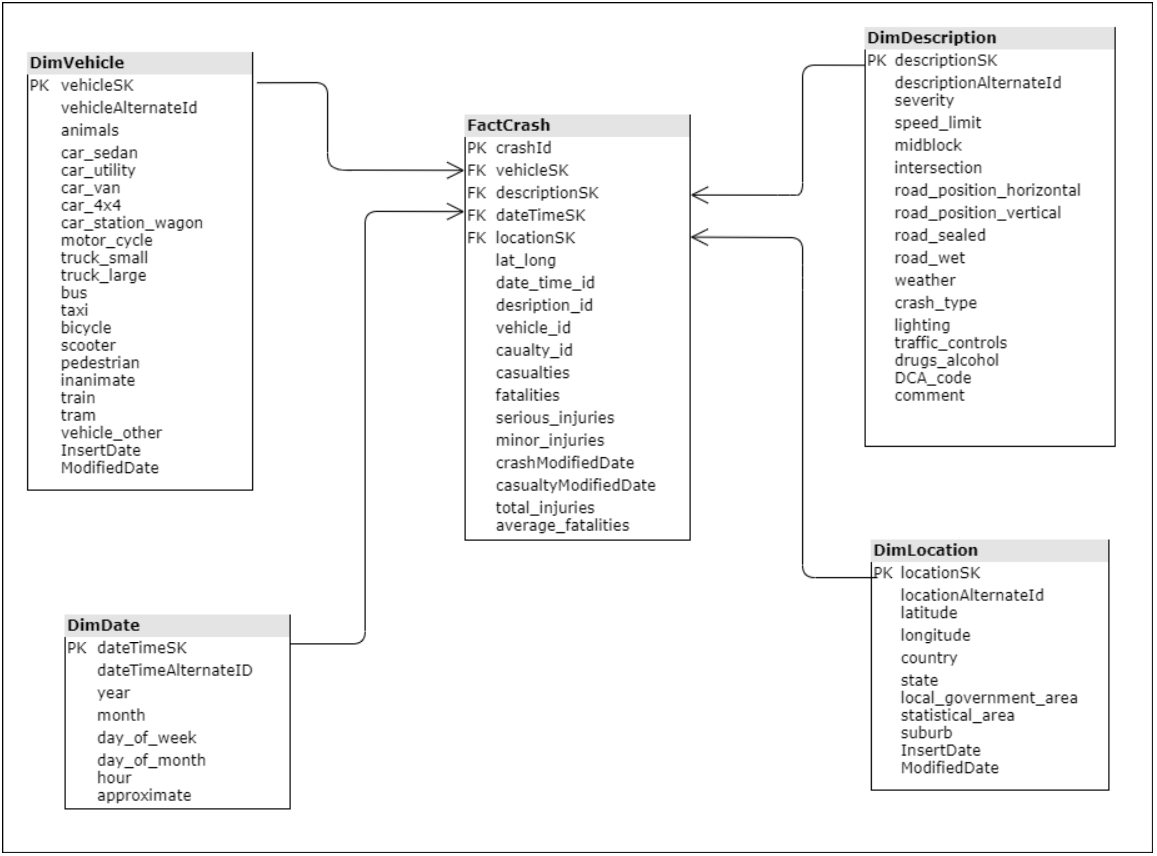


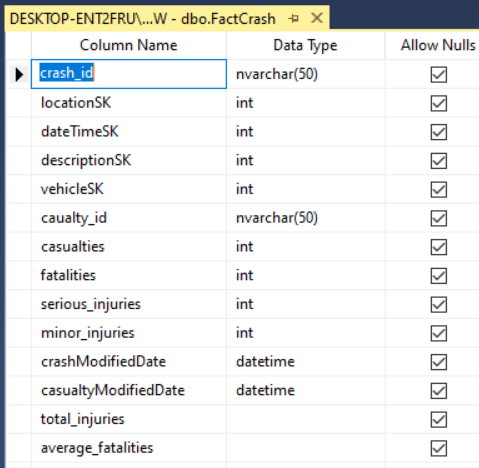
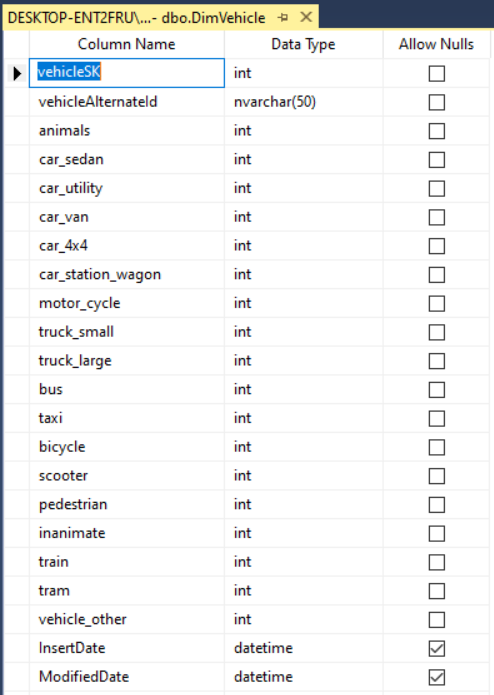
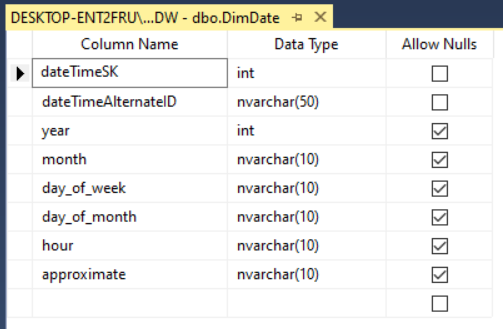


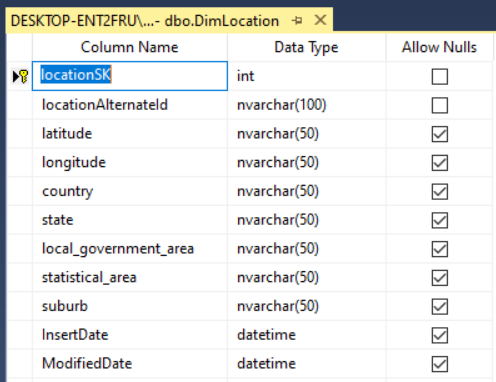


**Solution Architecture**



**Data warehouse design & development**

**Data Warehouse Data types**



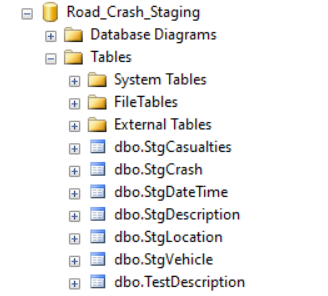
I have implemented a star schema where I have used 1 fact table – FactCrash and 4 DimensionTables where DimDescription was considered as Slowly Changing Dimension.

**Assumptions**

I have combined casualties table with the crash table when transforming data to data warehouse whereas I have considered Dimdescription table as a slowly dimension table as I have considered road\_position\_horizontal, road\_position\_vertical, road\_wet, road\_sealed comment as slowly changing attributes as in order to keep tracks on the conditions and the status of accident occurrence it is important to keep track of the history of this data.

**ETL Development**

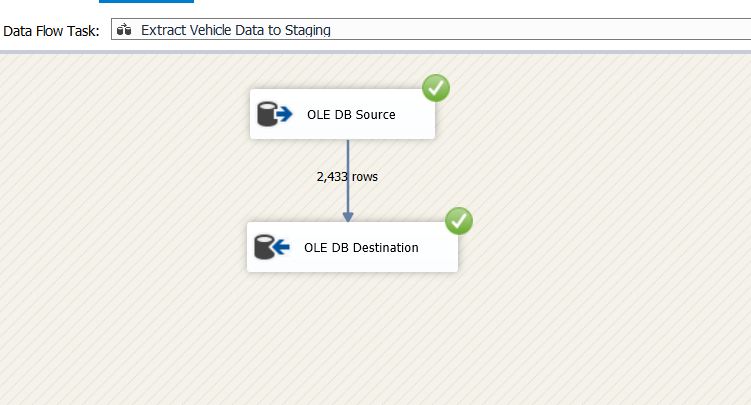
First of all, in order to extract all the data from tables to Staging to Road\_Crash\_Staging, I have used SSIS as belows



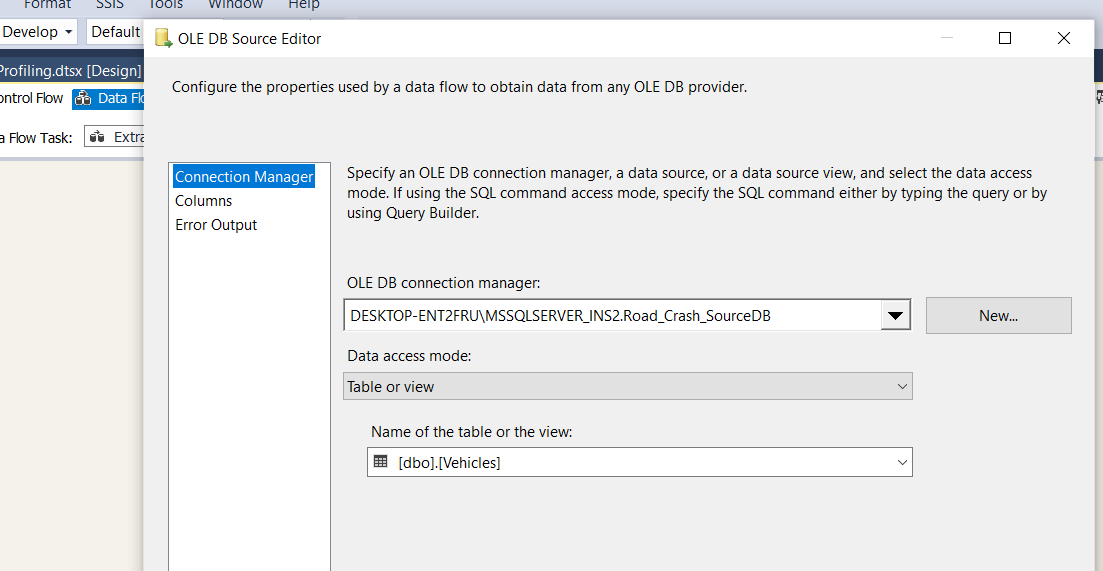


Below I have showed the steps I used to extract data to Staging from SSIS

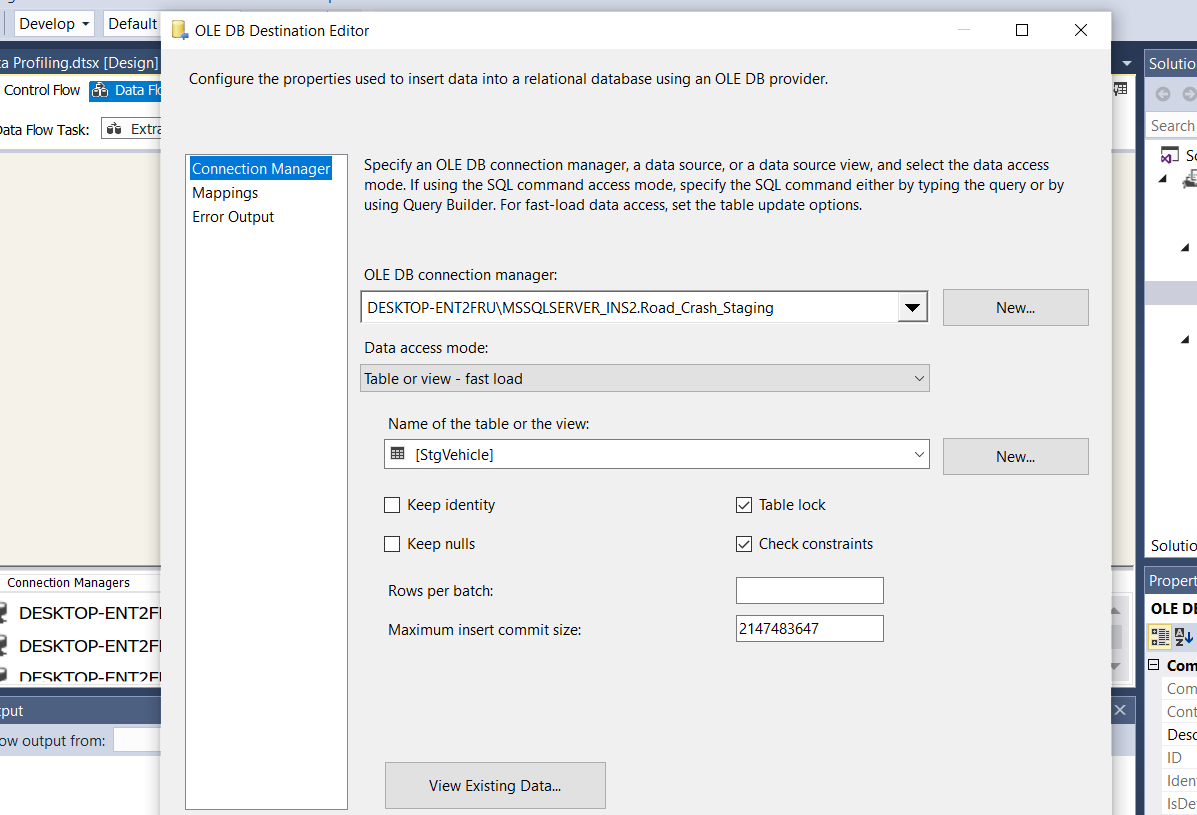
1. I have used OLE DB Source and OLE DB Destination as follows.



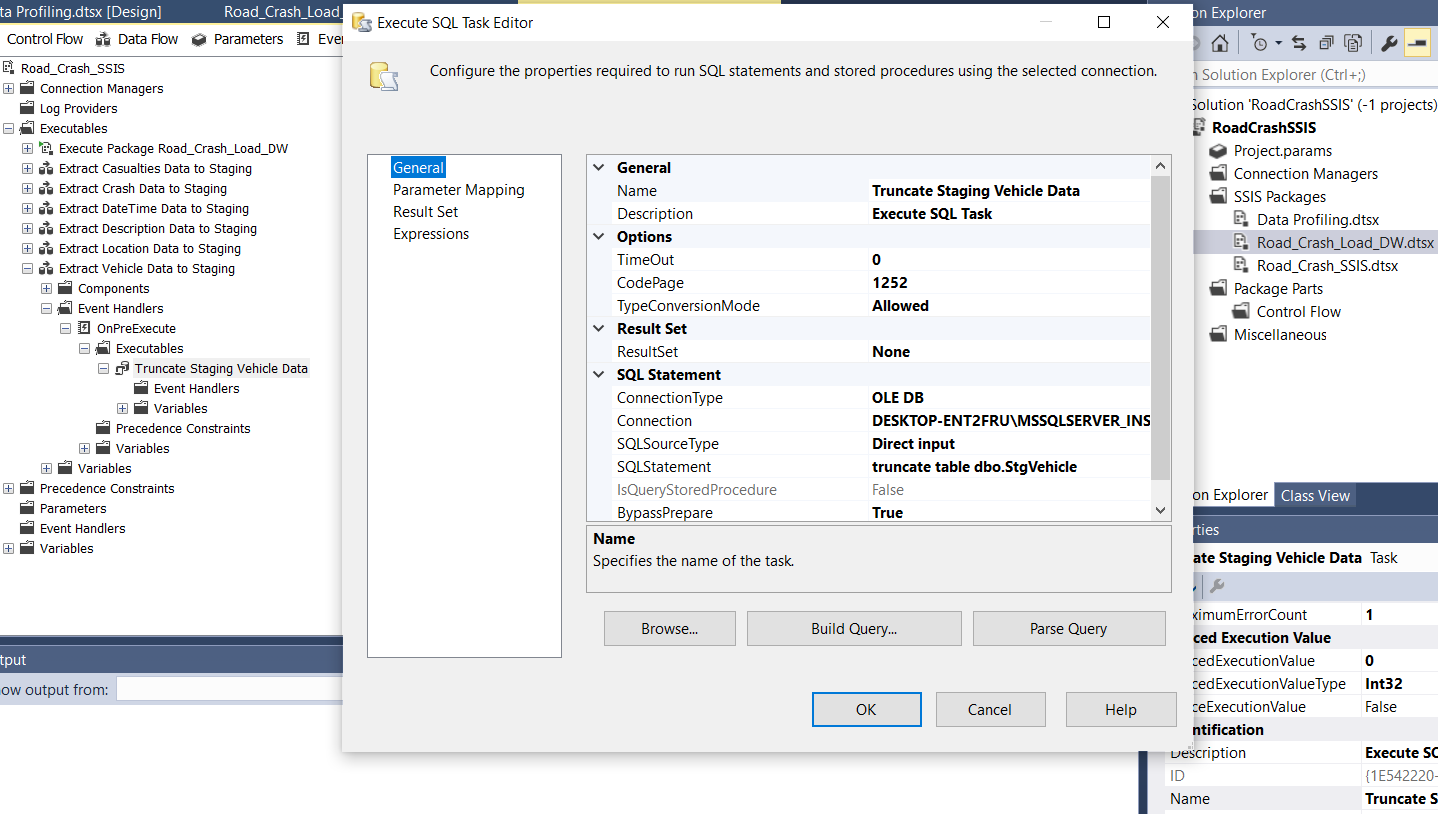
2. To extract data from Source

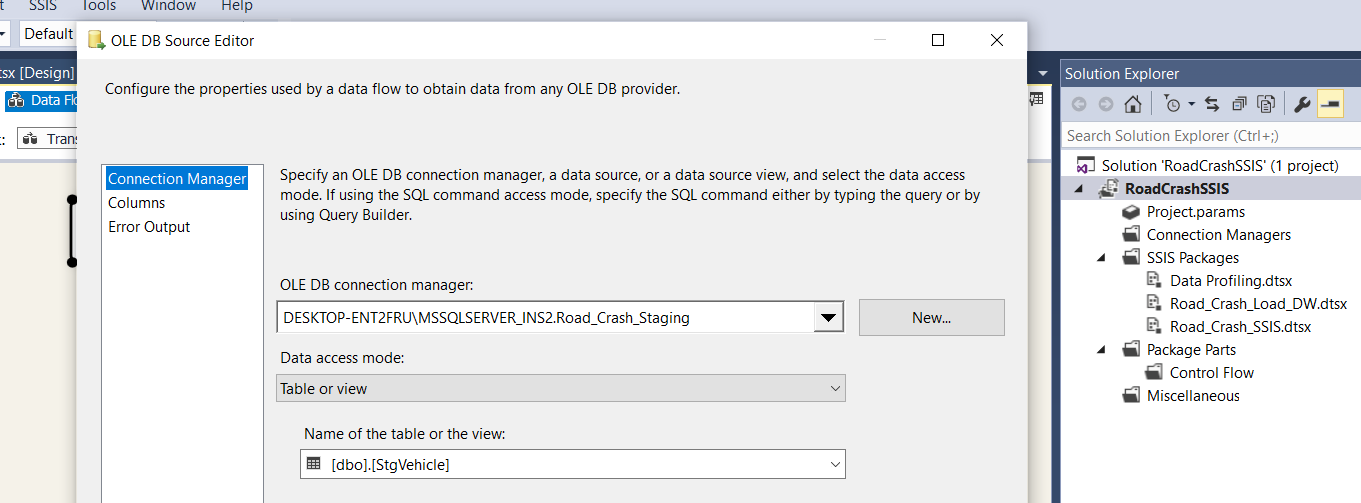


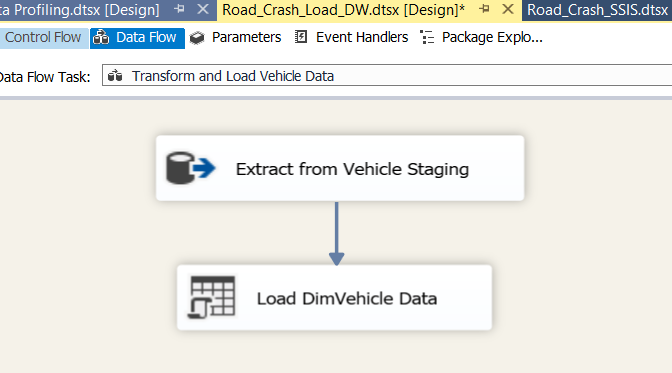
3. To load data to destination

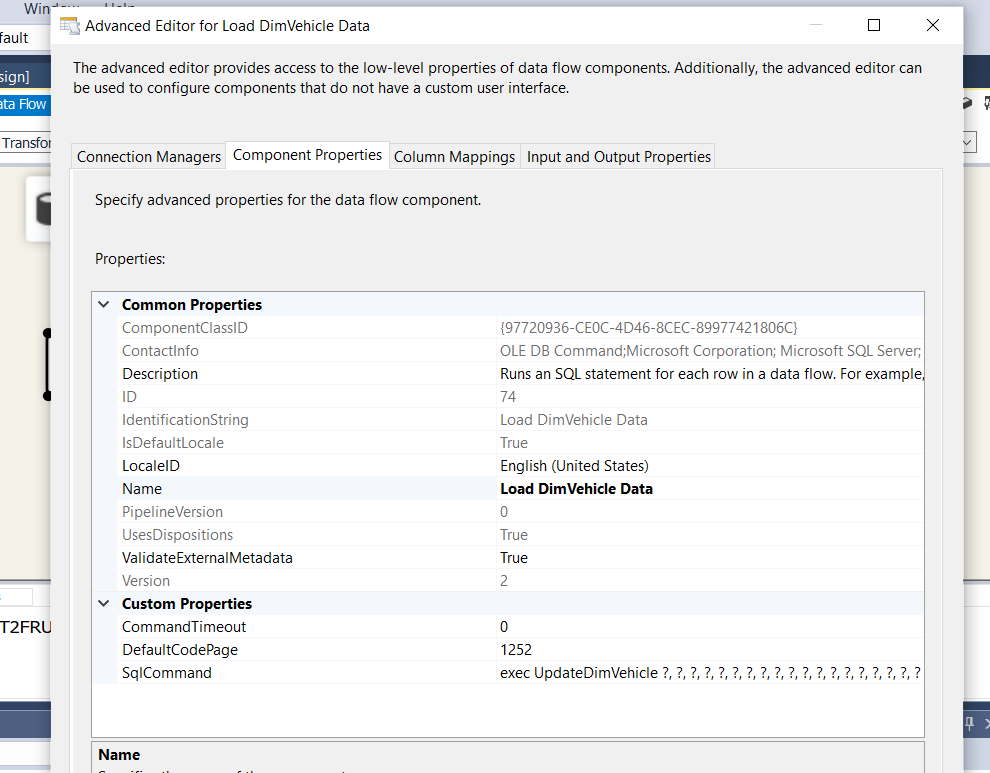


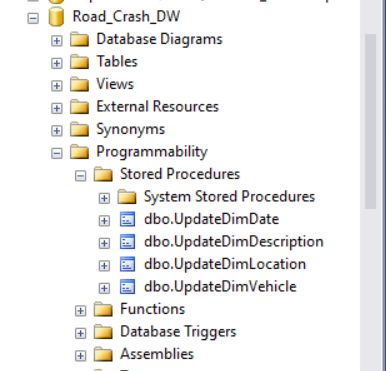
All Event Handlers I have used

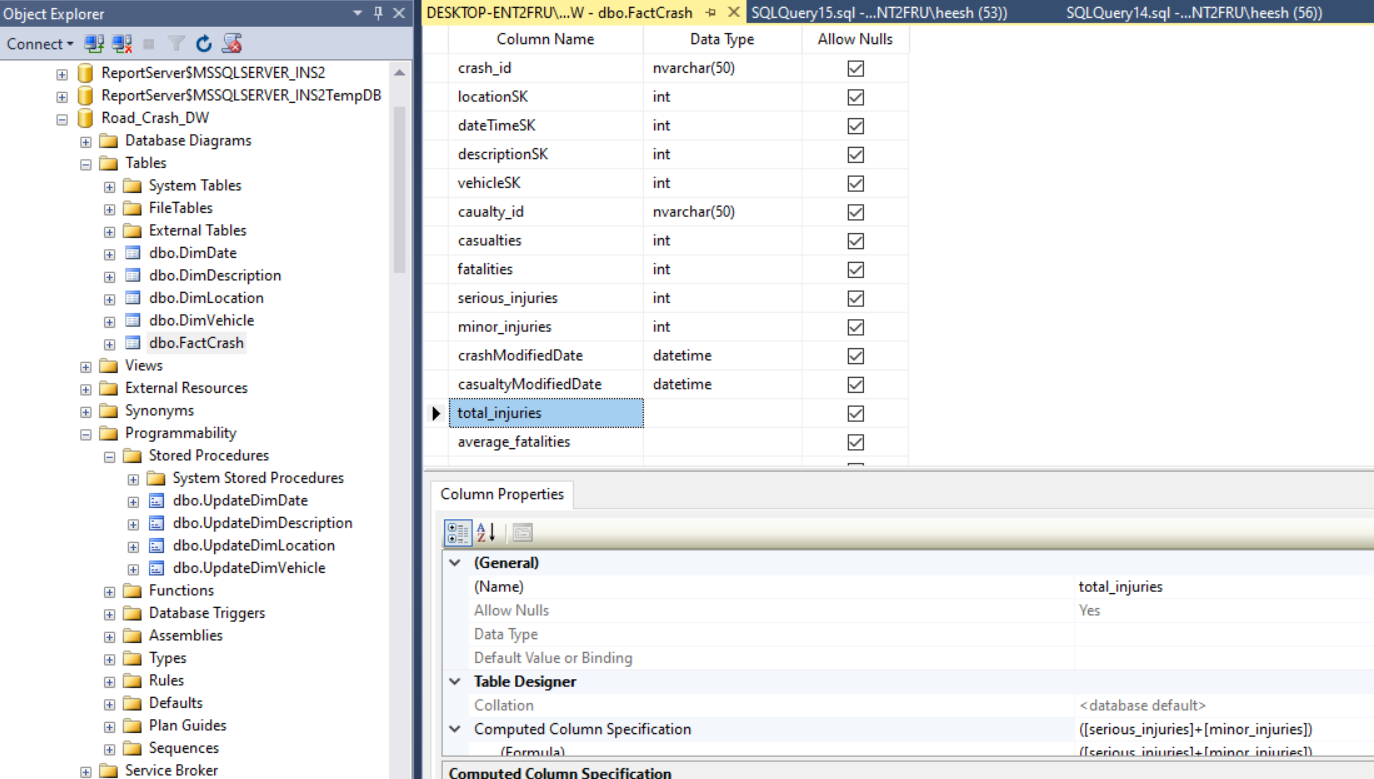


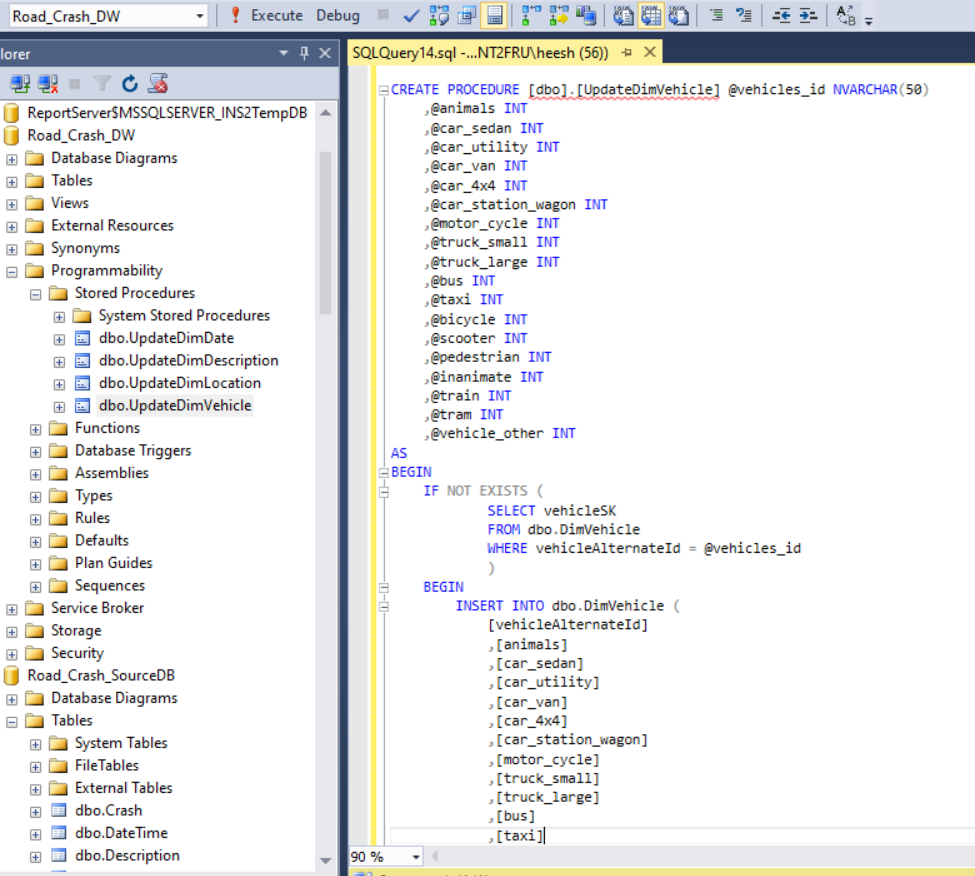


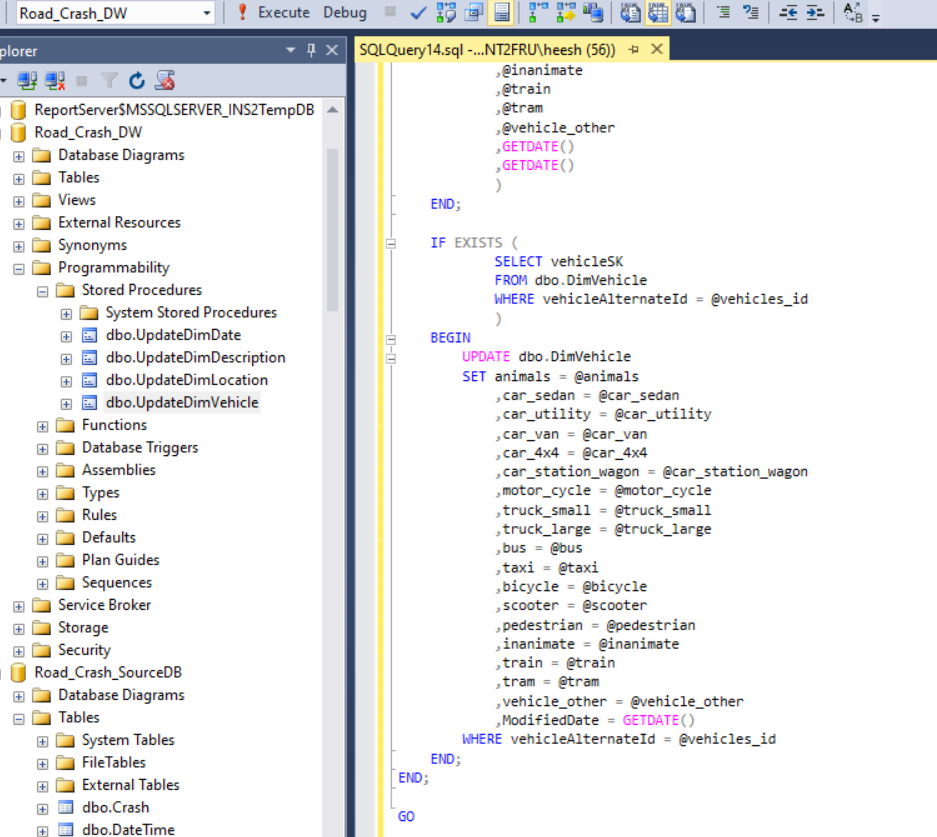




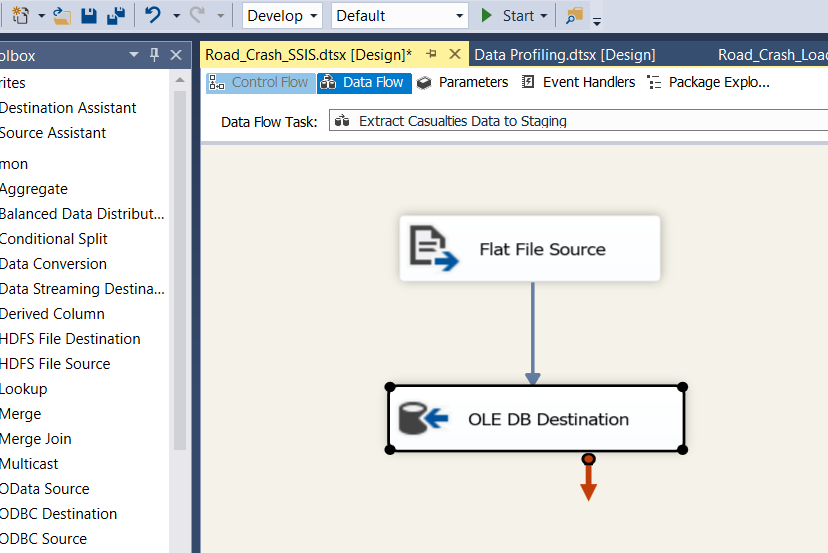


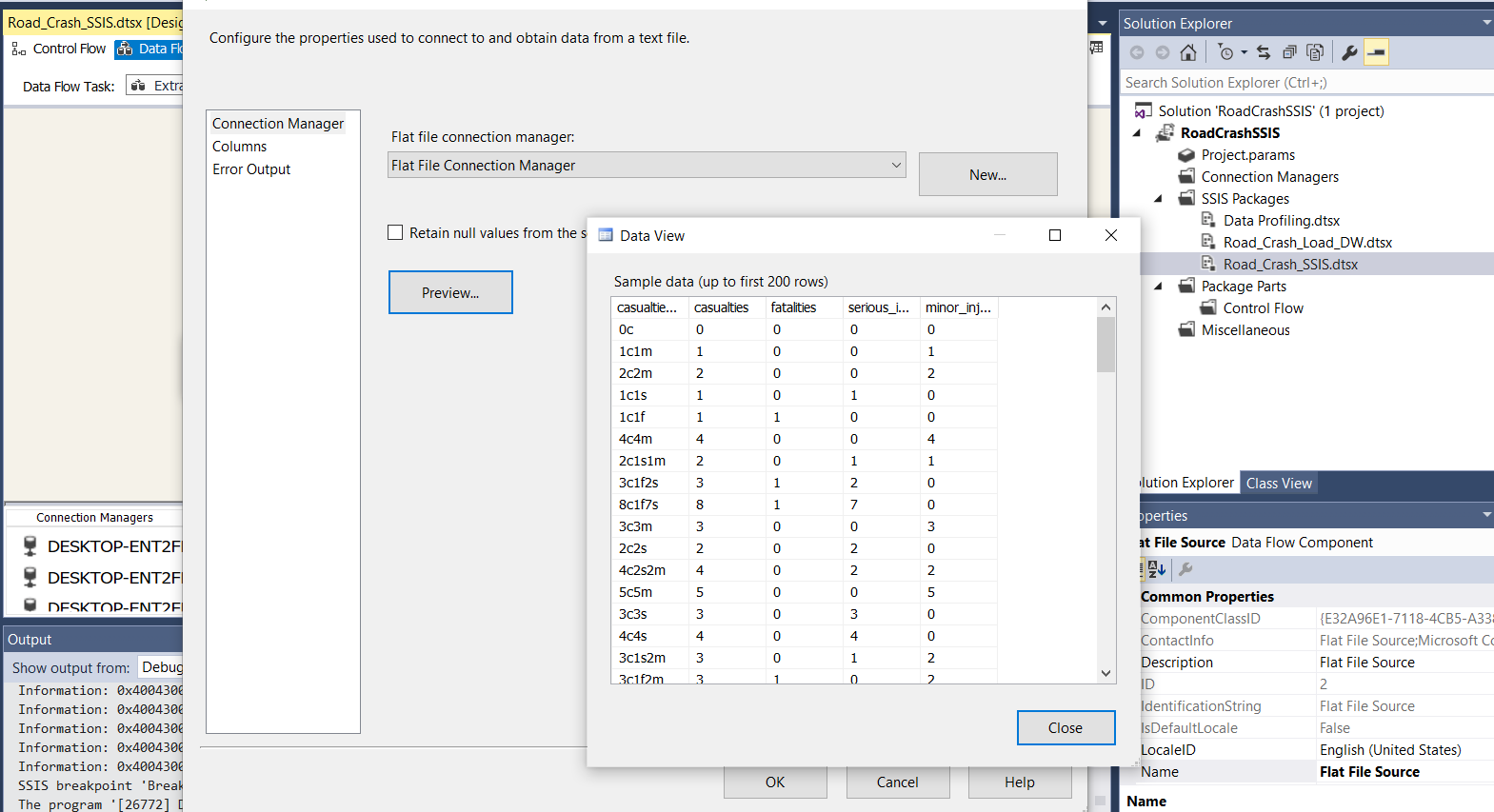


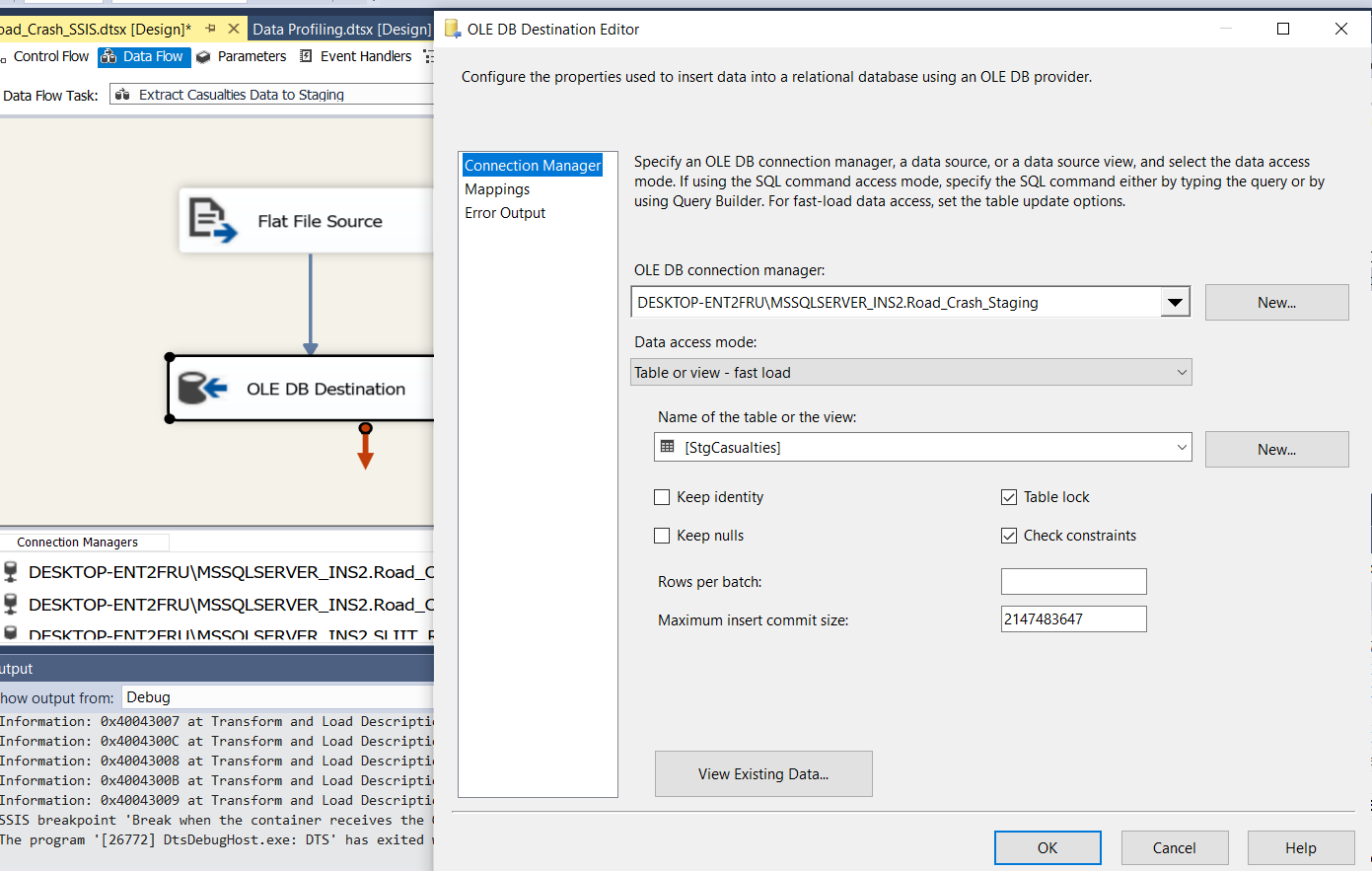




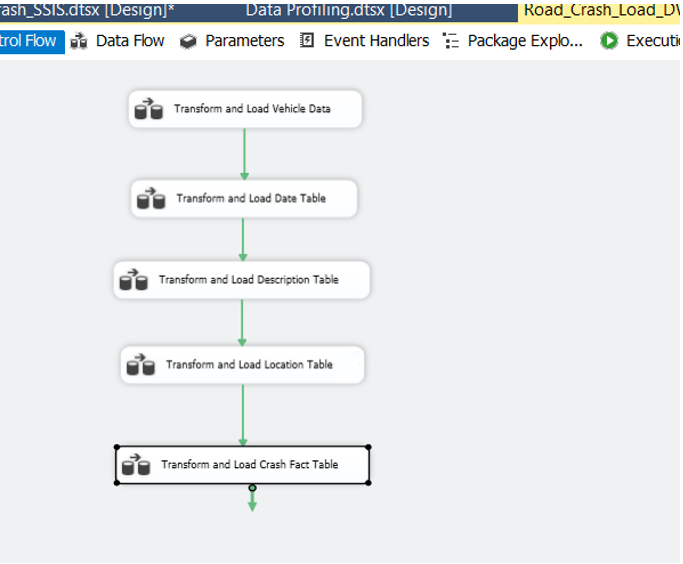
This is how I extracted data from the text file to Staging



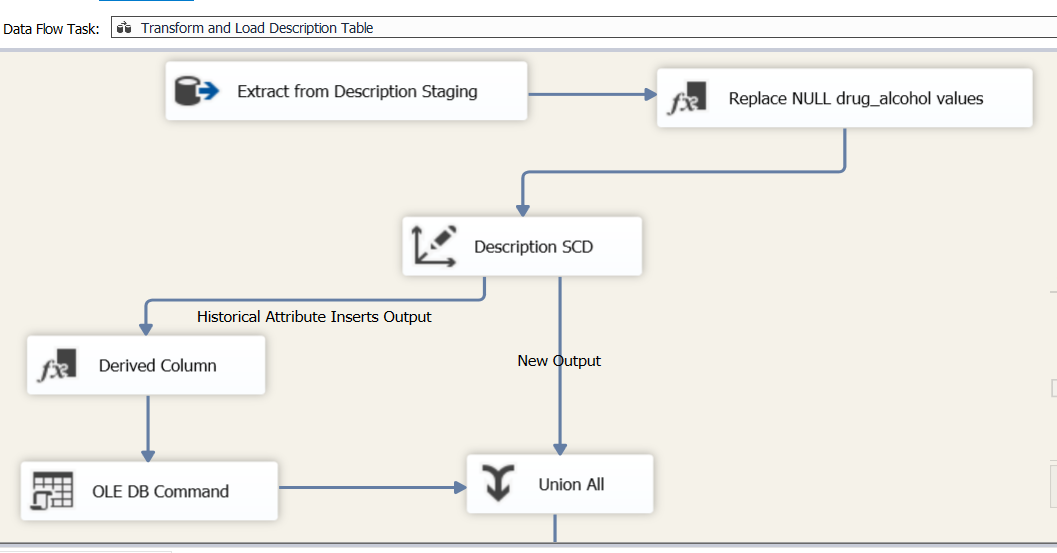


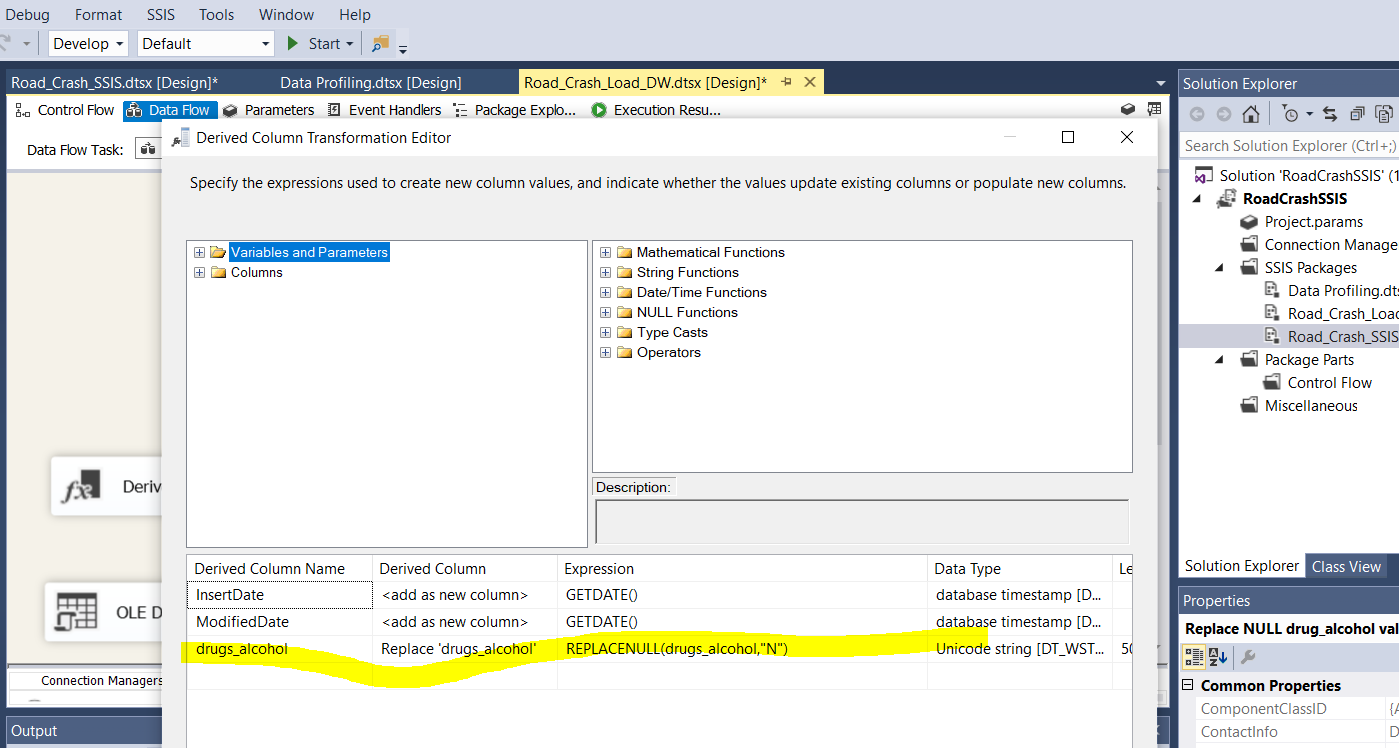


Extracting data from Staging to Data Warehouse

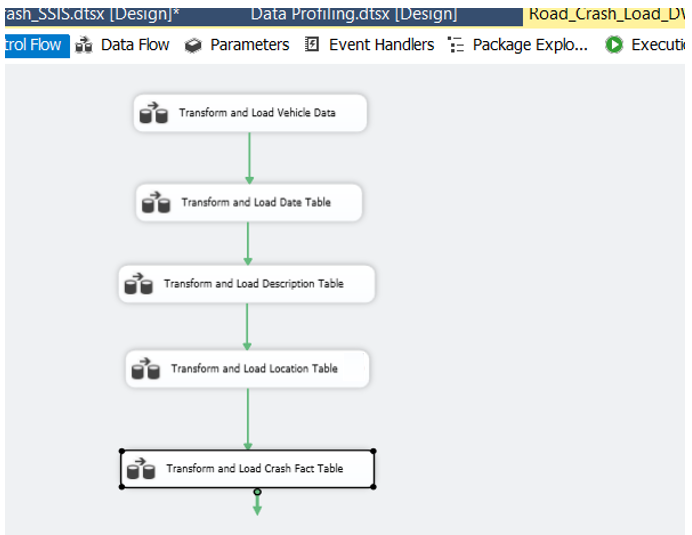


My use of slowly changing dimension on DimDescription. Also I have used derived column in order to replace Null values with N for drugs\_alcohol field in DimDescription

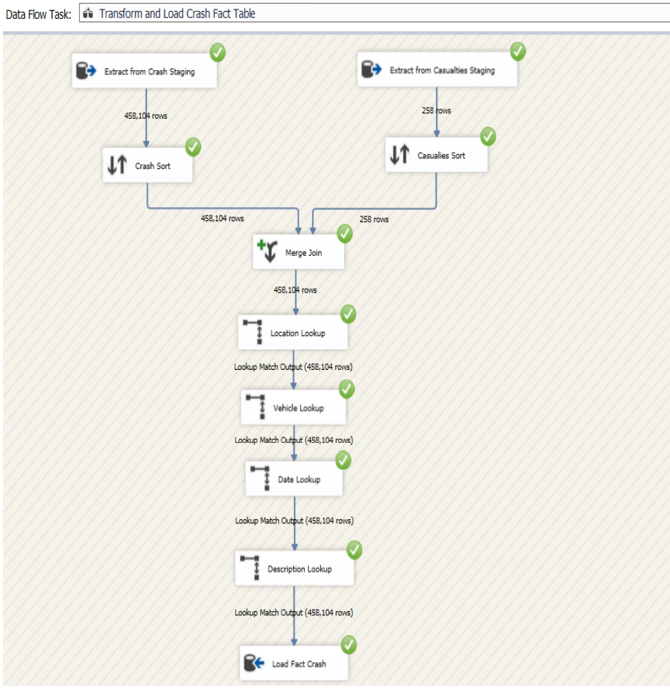




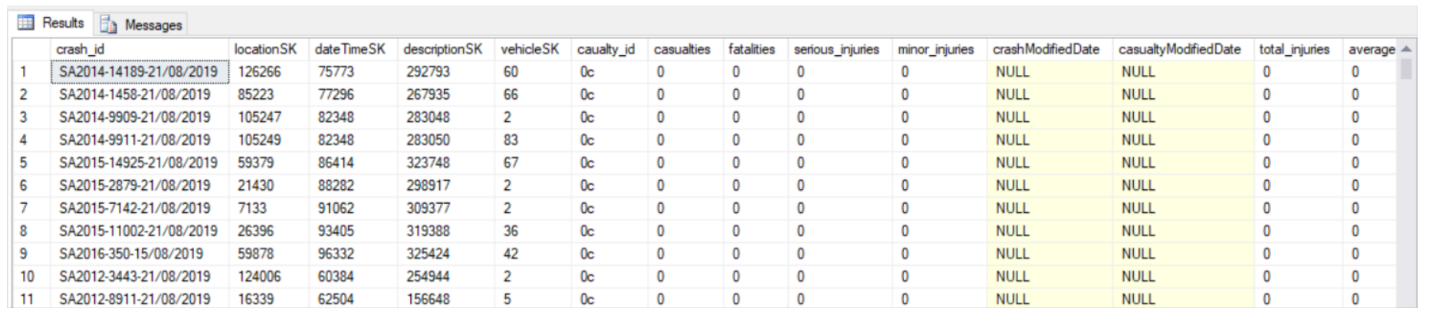


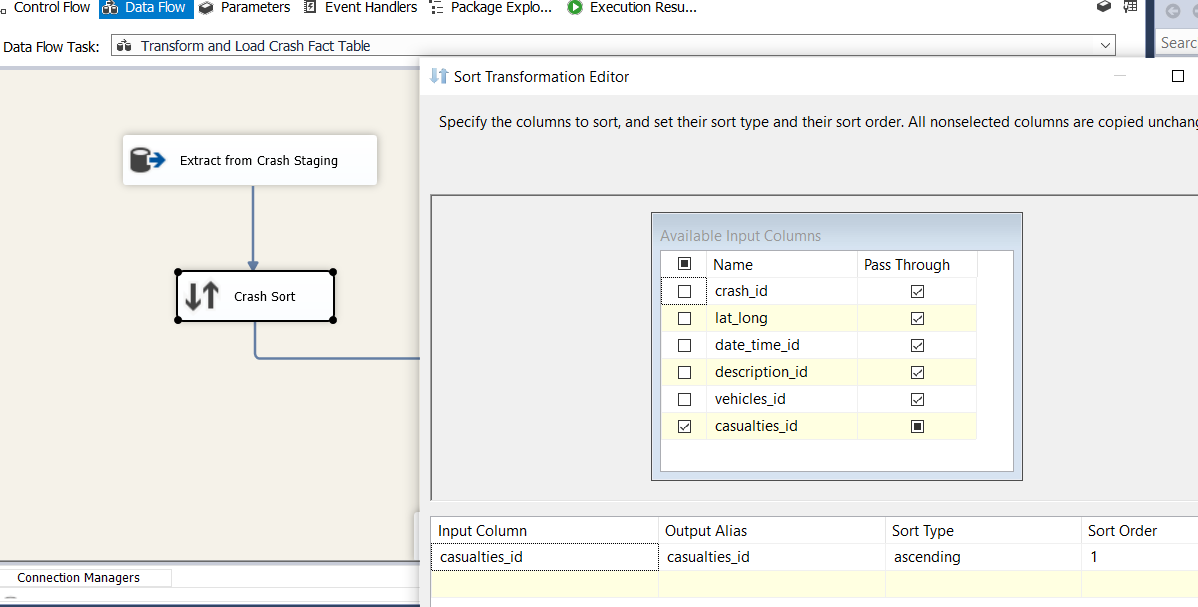


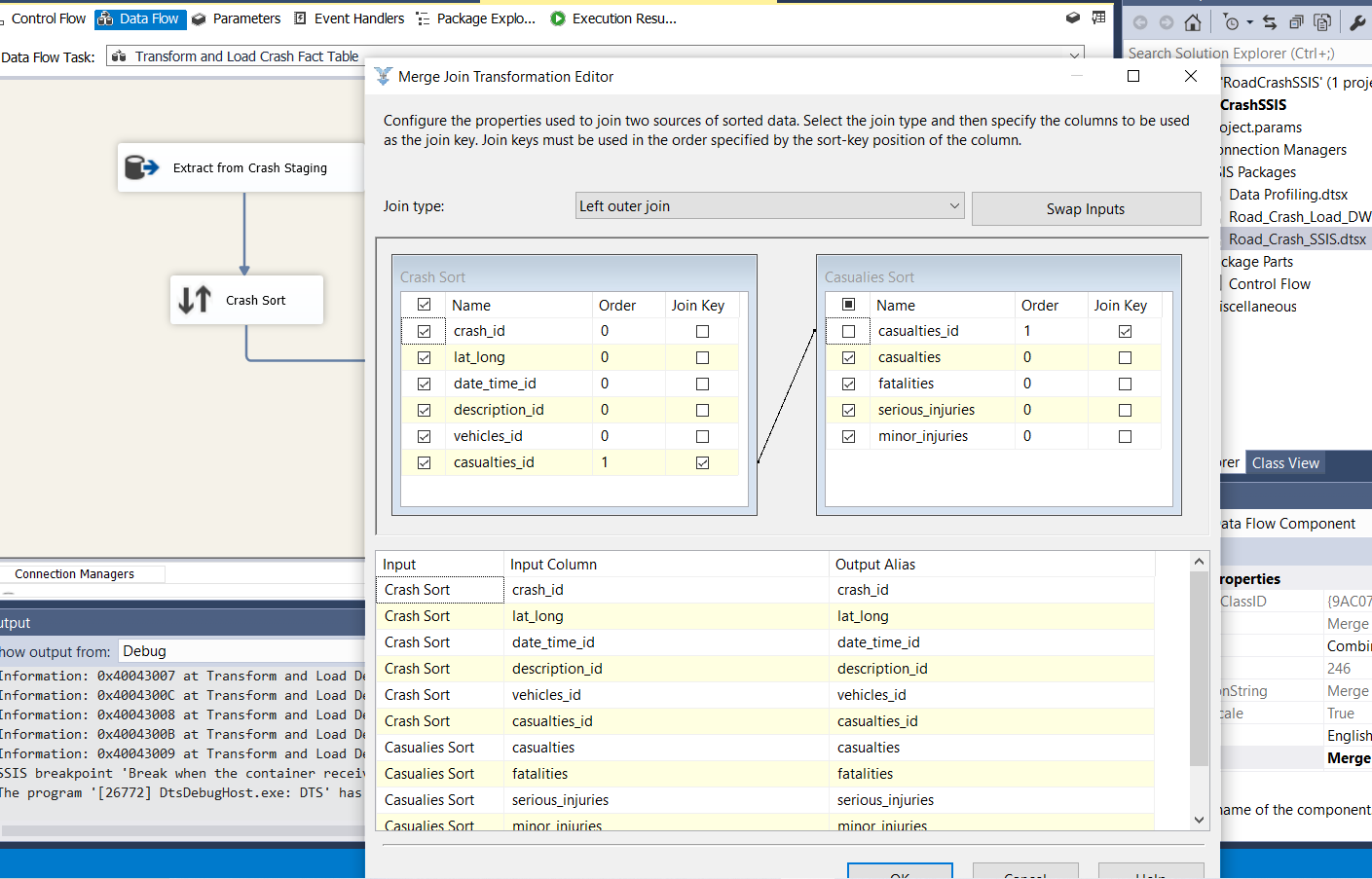
Below I have merged the extracted crash table with casualties table with crash table sorting them with casualties\_id and then I have used for lookups in order to load the transformed data into data warehouse with the use of surrogate keys

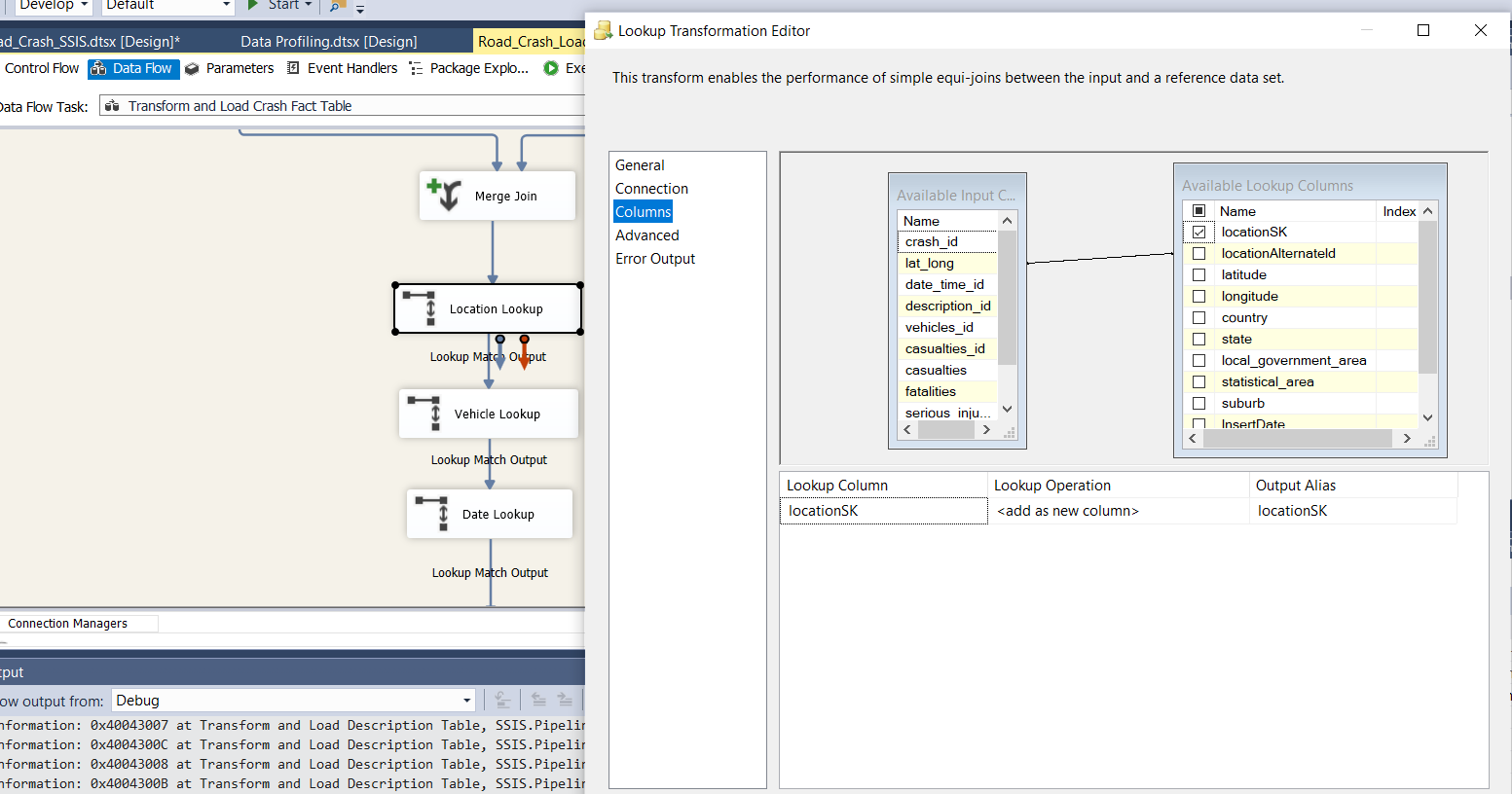


Below is Fact Crash Table in Data Warehouse









Data Profiling

