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----------

(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.

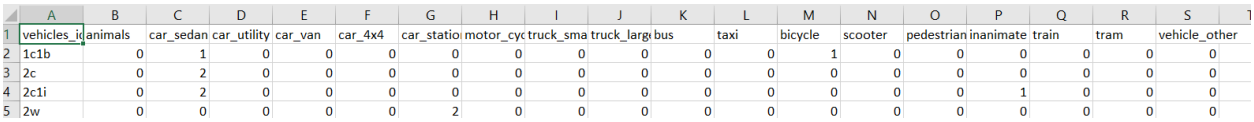
ER Diagram

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

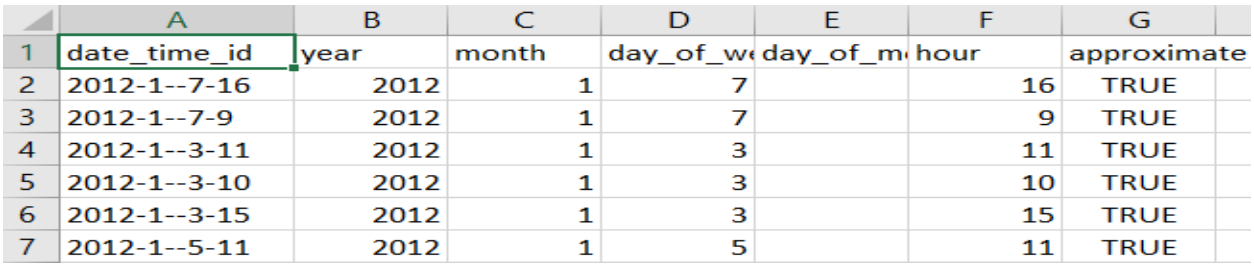
Vehicle.csv



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	vehicles_id	animals	car_sedan	car_utility	car_van	car_4x4	car_statio	motor_cyc	truck_sma	truck_larg	bus	taxi	bicycle	scooter	pedestrian	inanimate	train	tram	vehicle_other
2	1c1b	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
3	2c	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	2c1i	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
5	2w	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0

Location.csv

DateTime.csv



	A	B	C	D	E	F	G
1	date_time_id	year	month	day_of_we	day_of_m	hour	approximate
2	2012-1--7-16	2012	1	7		16	TRUE
3	2012-1--7-9	2012	1	7		9	TRUE
4	2012-1--3-11	2012	1	3		11	TRUE
5	2012-1--3-10	2012	1	3		10	TRUE
6	2012-1--3-15	2012	1	3		15	TRUE
7	2012-1--5-11	2012	1	5		11	TRUE

Crash.csv

Casualties.csv

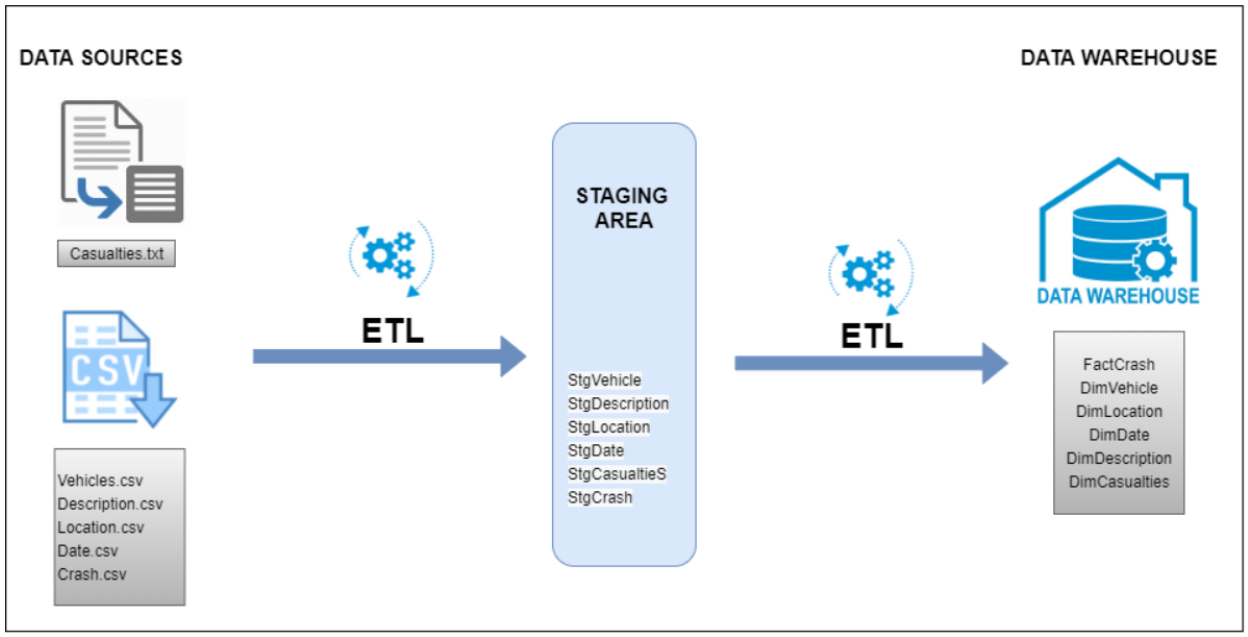
Casualties.txt

Data types of source tables

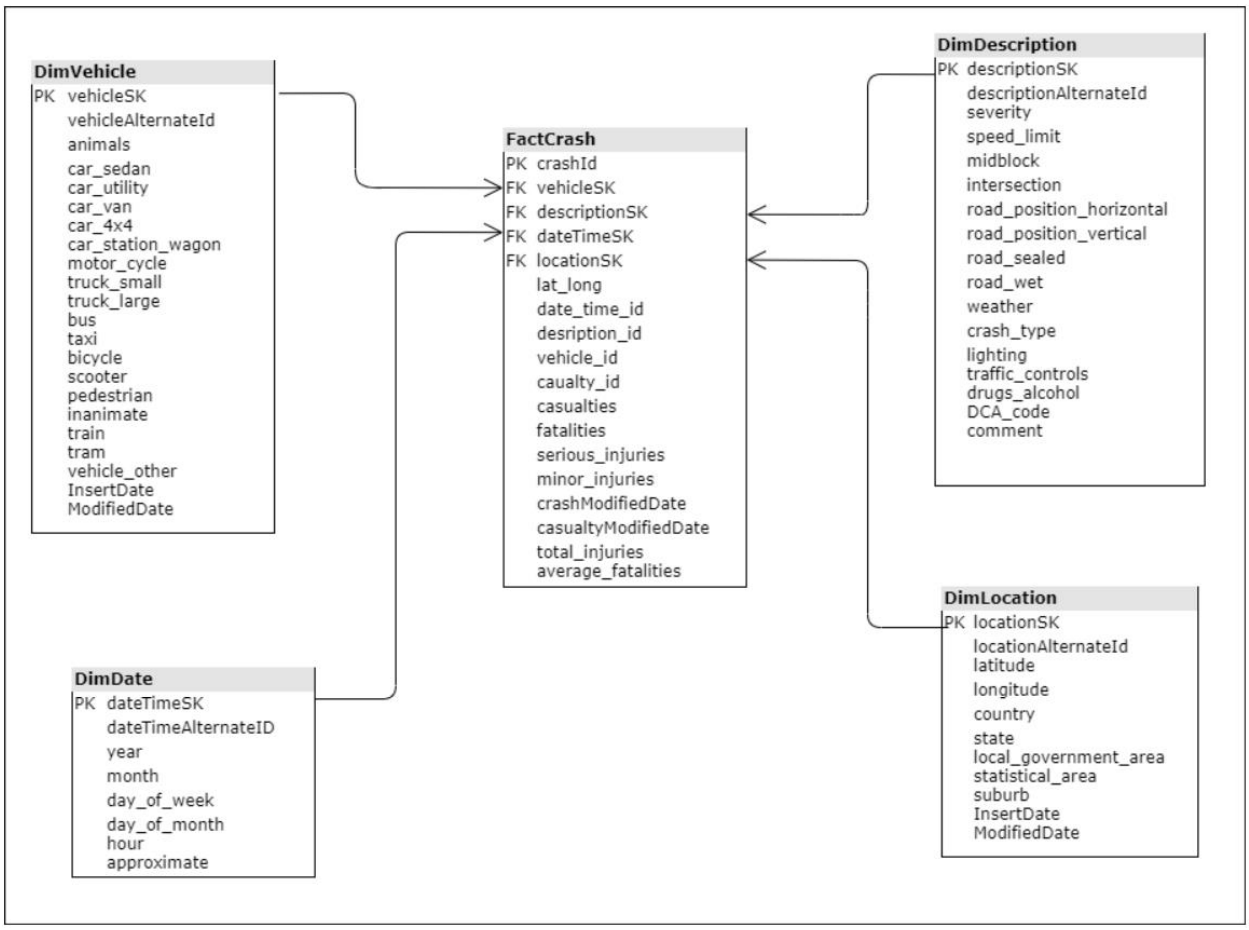
I have created Road_Crash_SourceDB Database ad imported the CSV files of source data.

Column Name	Data Type	Allow Nulls
date_time_id	nvarchar(50)	☐
year	int	☑
month	nvarchar(10)	☑
day_of_week	nvarchar(10)	☑
day_of_month	nvarchar(10)	☑
hour	nvarchar(10)	☑
approximate	nvarchar(10)	☑

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

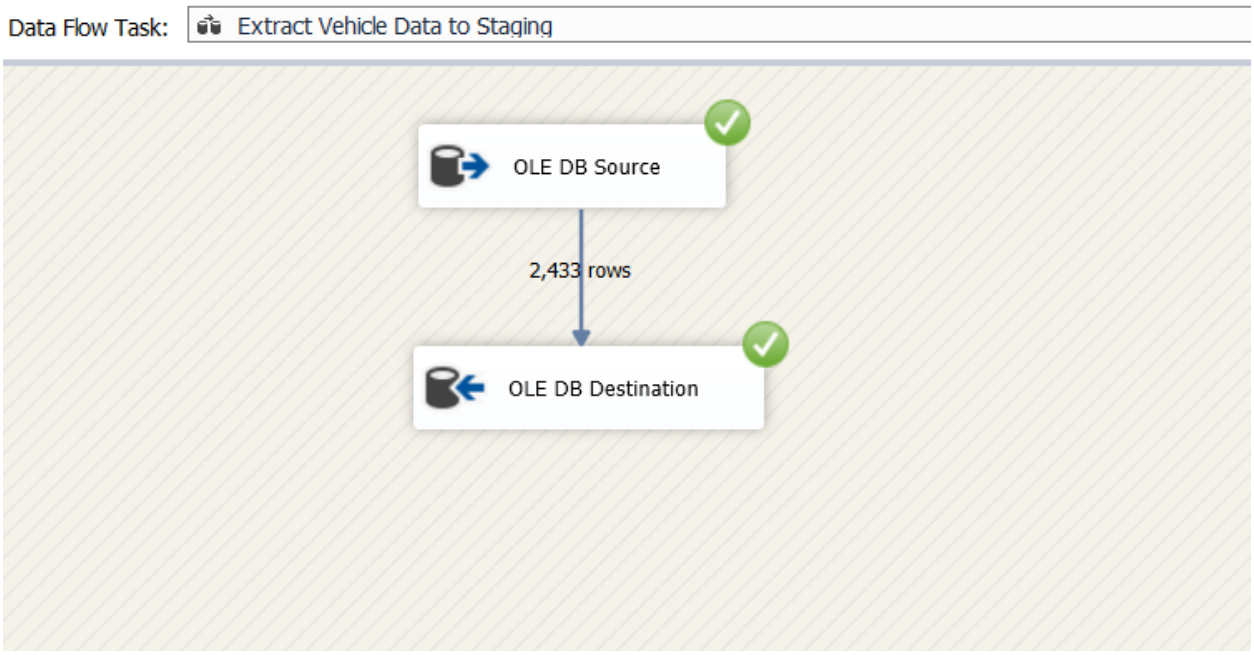
I have combined casualties table with the crash table when transforming data to data warehouse whereas I have considered Dimdescription table as a slowly dimension table as I have considered road_position_horizontal, road_position_vertical, road_wet, road_sealed comment as slowly changing attributes 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

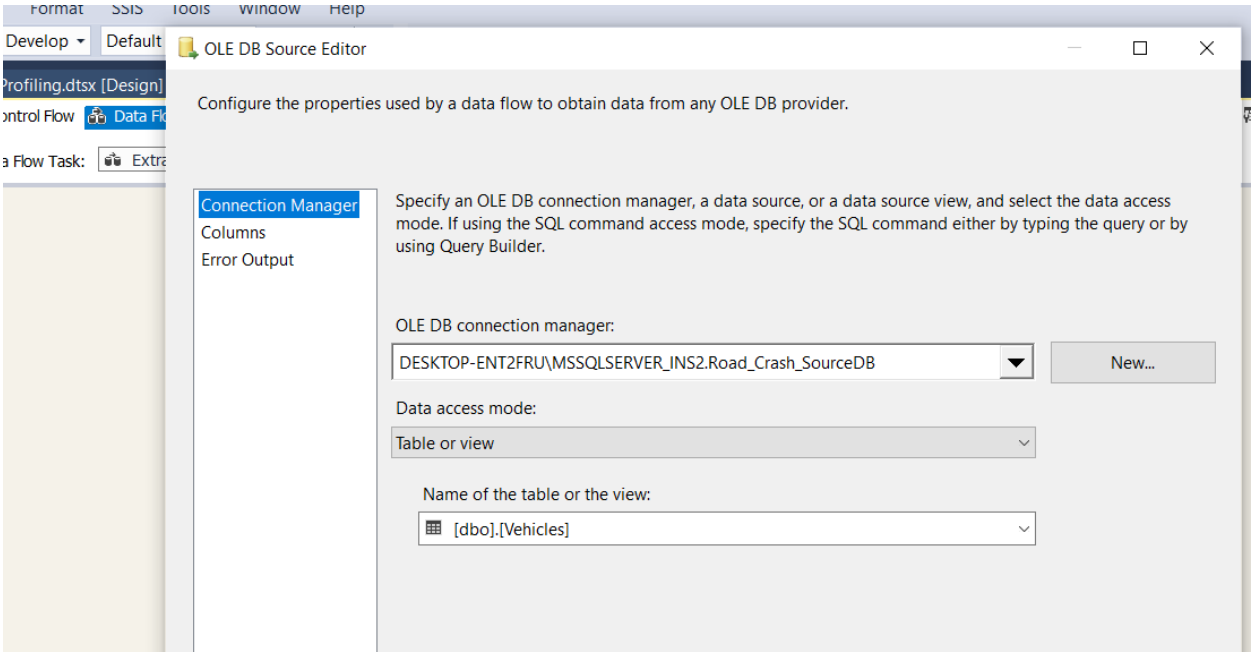
First of all, in order to extract all the data from tables to Staging to Road_Crash_Staging, I have used SSIS as belows

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

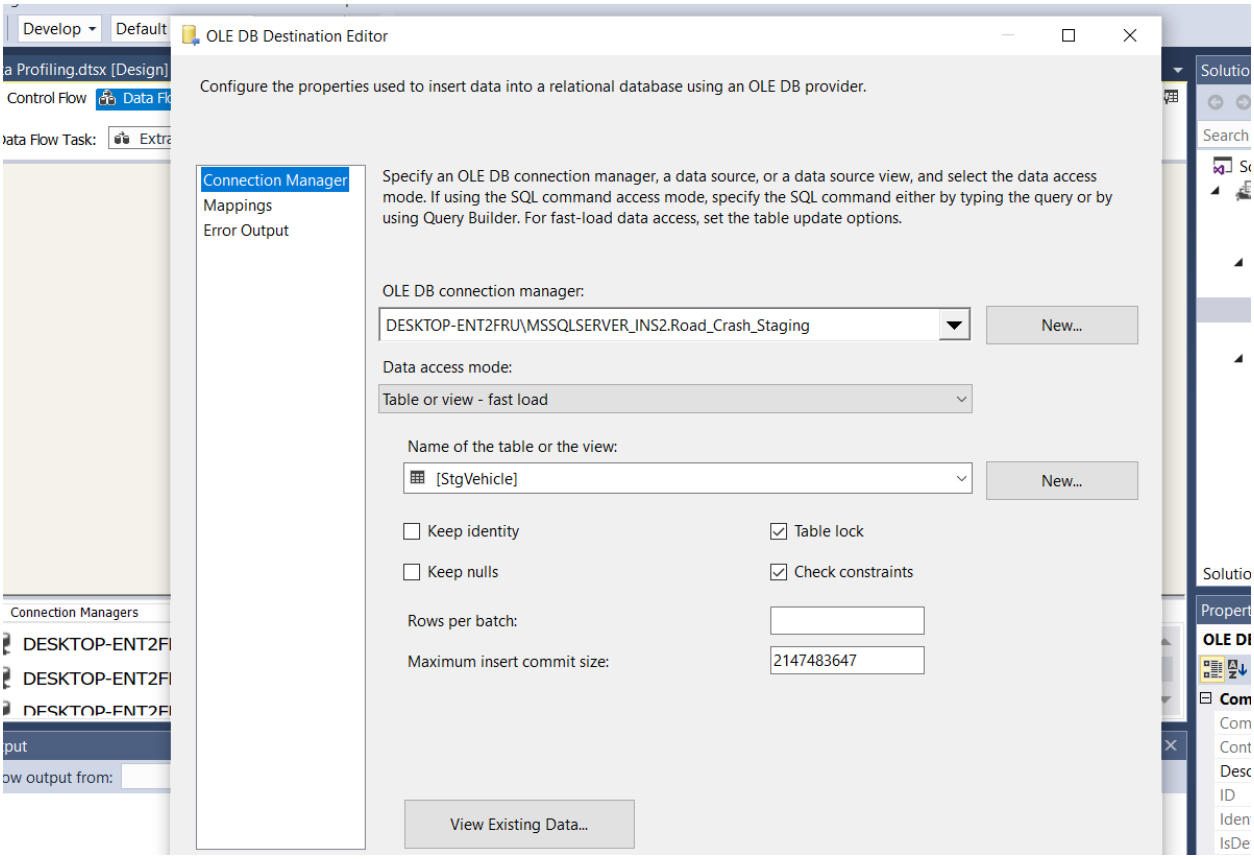
1. I have used OLE DB Source and OLE DB Destination as follows.



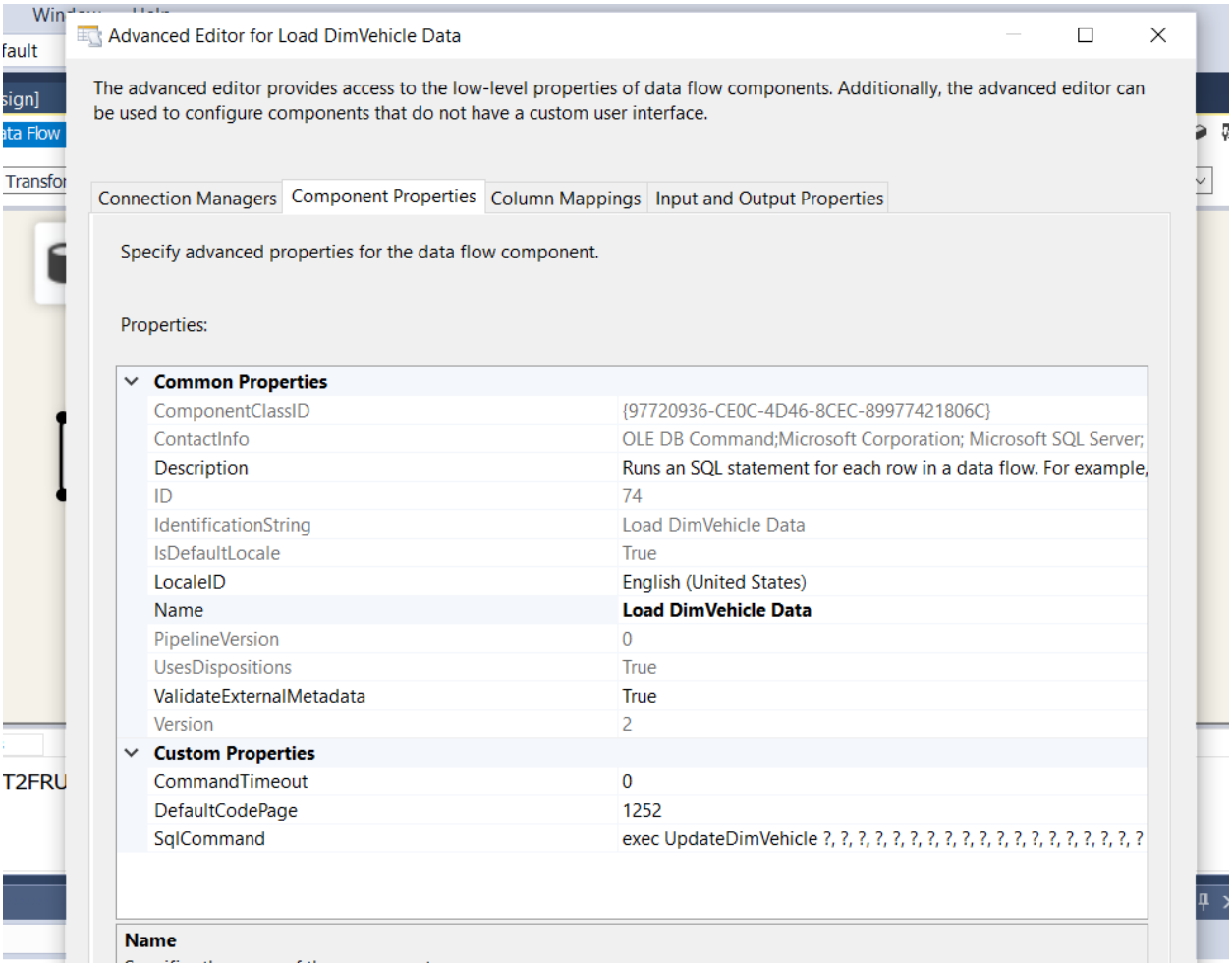
2. To extract data from Source

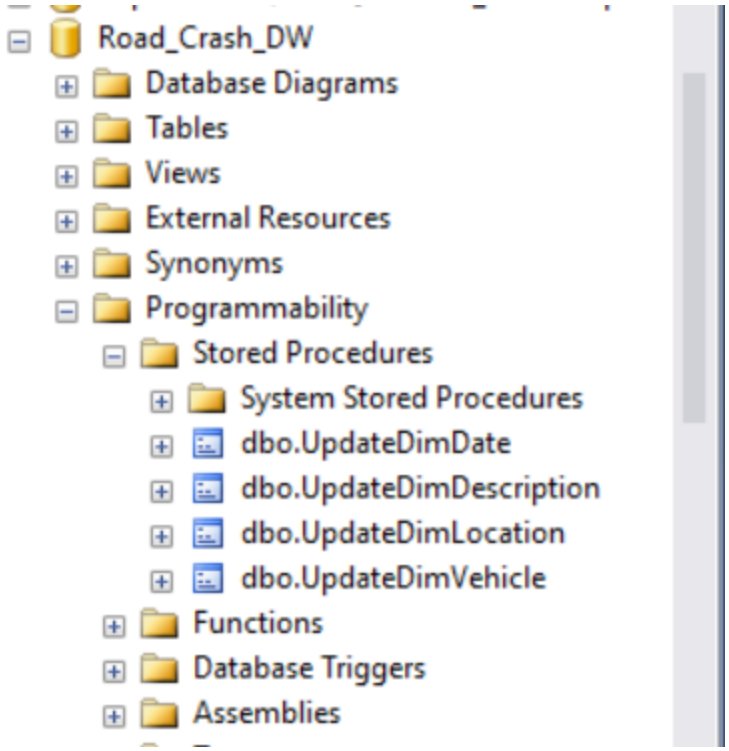


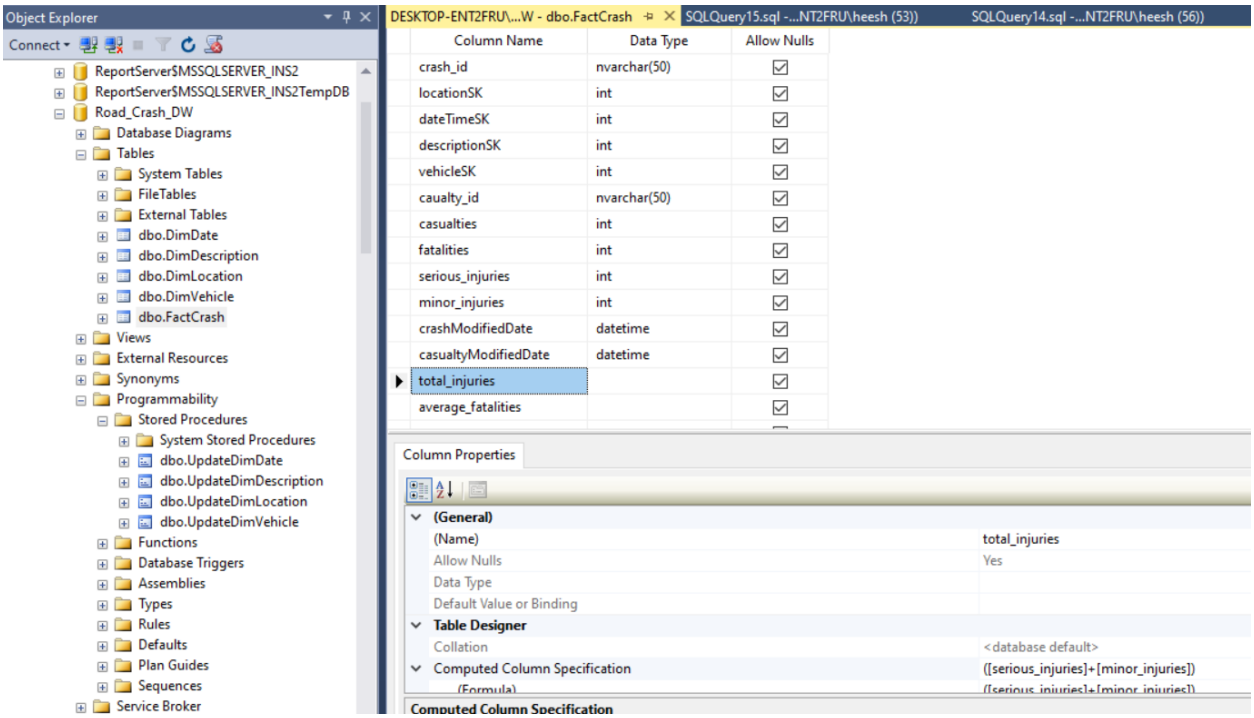
3. To load data to destination

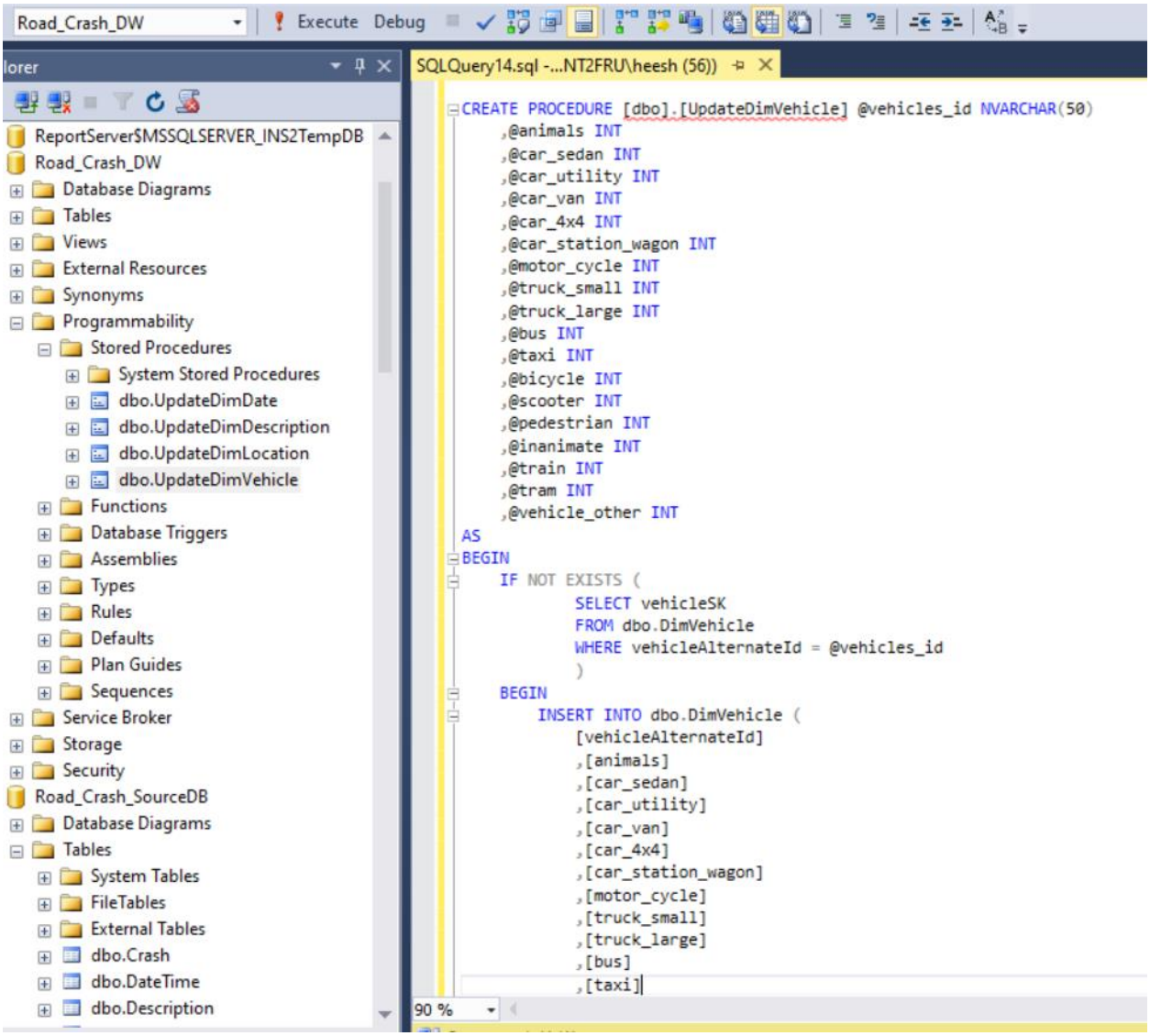


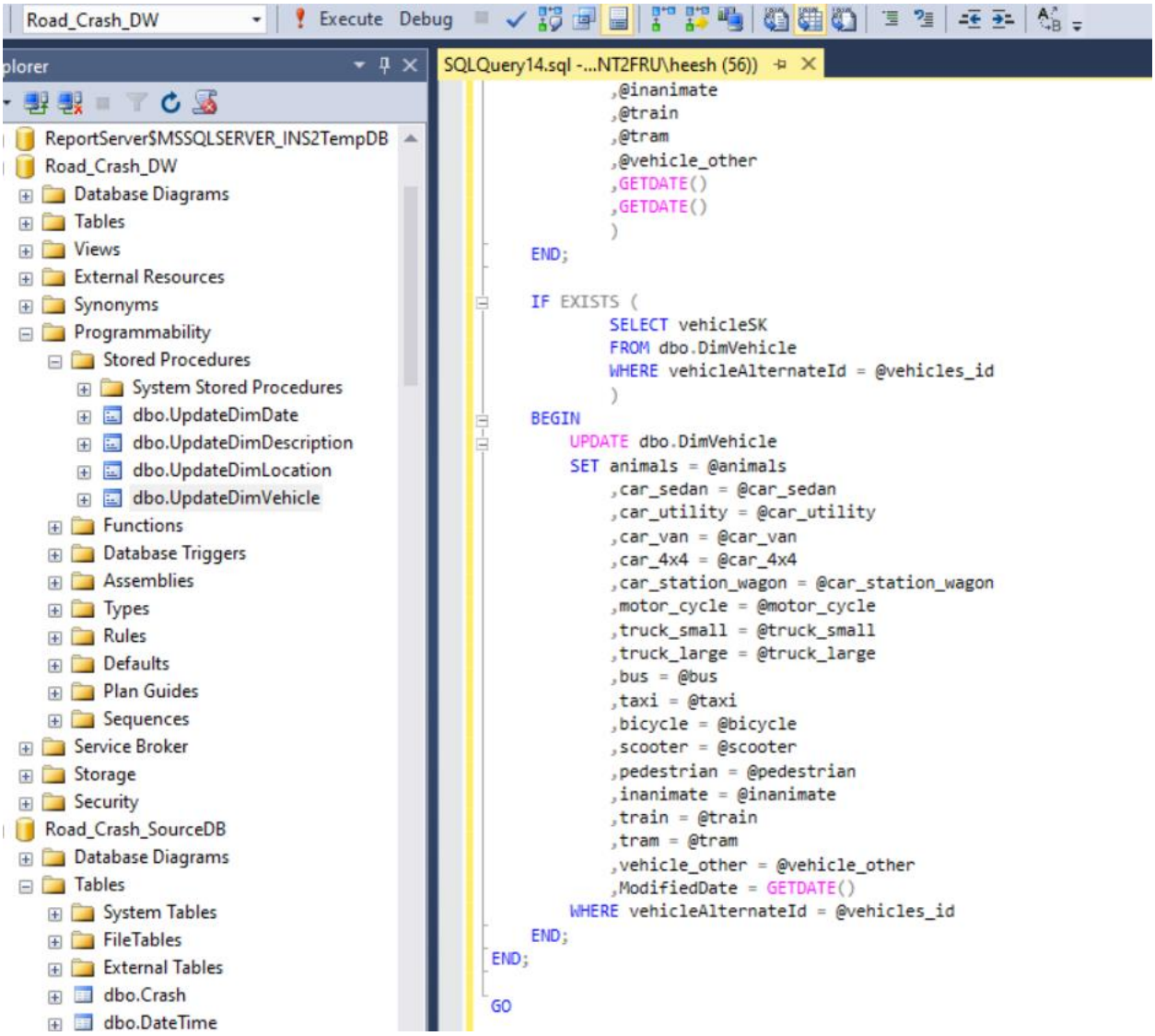
All Event Handlers I have used


Advanced Editor for Load DimVehicle Data
The advanced editor provides access to the low-level properties of data flow components. Additionally, the advanced editor can be used to configure components that do not have a custom user interface.
Connection Managers
Component Properties
Column Mappings
Input and Output Properties
Specify advanced properties for the data flow component.
Properties:
Common Properties
ComponentClassID {97720936-CE0C-4D46-8CEC-89977421806C}
ContactInfo OLE DB Command;Microsoft Corporation; Microsoft SQL Server;
Description Runs an SQL statement for each row in a data flow. For example,
ID 74
IdentificationString Load DimVehicle Data
IsDefaultLocale True
LocaleID English (United States)
Name Load DimVehicle Data
PipelineVersion 0
UsesDispositions True
ValidateExternalMetadata True
Version 2
Custom Properties
CommandTimeout 0
DefaultCodePage 1252
SqlCommand exec UpdateDimVehicle ?, ?
Name

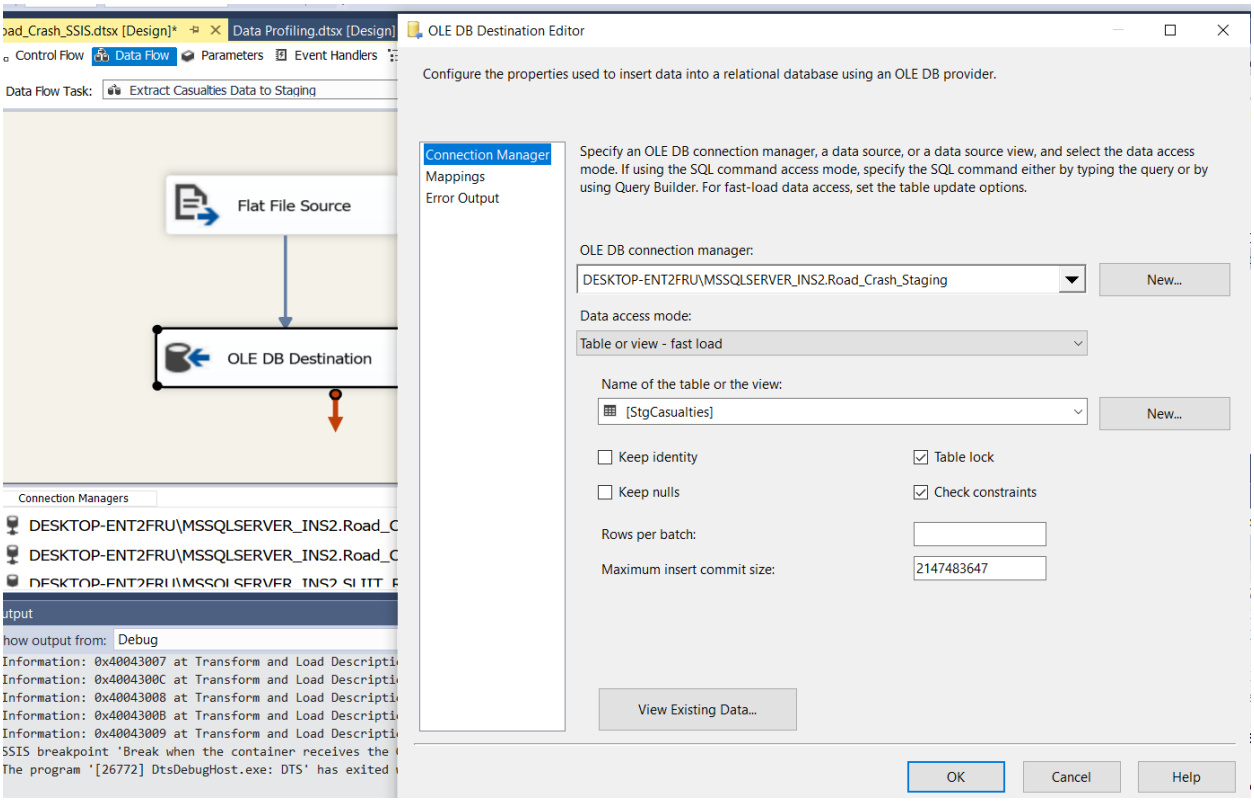

Road_Crash_DW
Database Diagrams
Tables
Views
External Resources
Synonyms
Programmability
Stored Procedures
System Stored Procedures
dbo.UpdateDimDate
dbo.UpdateDimDescription
dbo.UpdateDimLocation
dbo.UpdateDimVehicle
Functions
Database Triggers
Assemblies


Object Explorer
Connect
ReportServer$MSSQLSERVER_INS2
ReportServer$MSSQLSERVER_INS2TempDB
Road_Crash_DW
Database Diagrams
Tables
System Tables
FileTables
External Tables
dbo.DimDate
dbo.DimDescription
dbo.DimLocation
dbo.DimVehicle
dbo.FactCrash
Views
External Resources
Synonyms
Programmability
Stored Procedures
System Stored Procedures
dbo.UpdateDimDate
dbo.UpdateDimDescription
dbo.UpdateDimLocation
dbo.UpdateDimVehicle
Functions
Database Triggers
Assemblies
Types
Rules
Defaults
Plan Guides
Sequences
Service Broker
DESKTOP-ENT2FRU\...W - dbo.FactCrash
SQLQuery15.sql -...NT2FRU\heesh (53))
SQLQuery14.sql -...NT2FRU\heesh (56))
Column Name
Data Type
Allow Nulls
crash_id nvarchar(50)
locationSK int
dateTimeSK int
descriptionSK int
vehicleSK int
caualty_id nvarchar(50)
casualties int
fatalities int
serious_injuries int
minor_injuries int
crashModifiedDate datetime
casualtyModifiedDate datetime
total_injuries
average_fatalities
Column Properties
(General)
(Name) total_injuries
Allow Nulls Yes
Data Type
Default Value or Binding
Table Designer
Collation <database default>
Computed Column Specification ([serious_injuries]+[minor_injuries])
(Formula)
Computed Column Specification

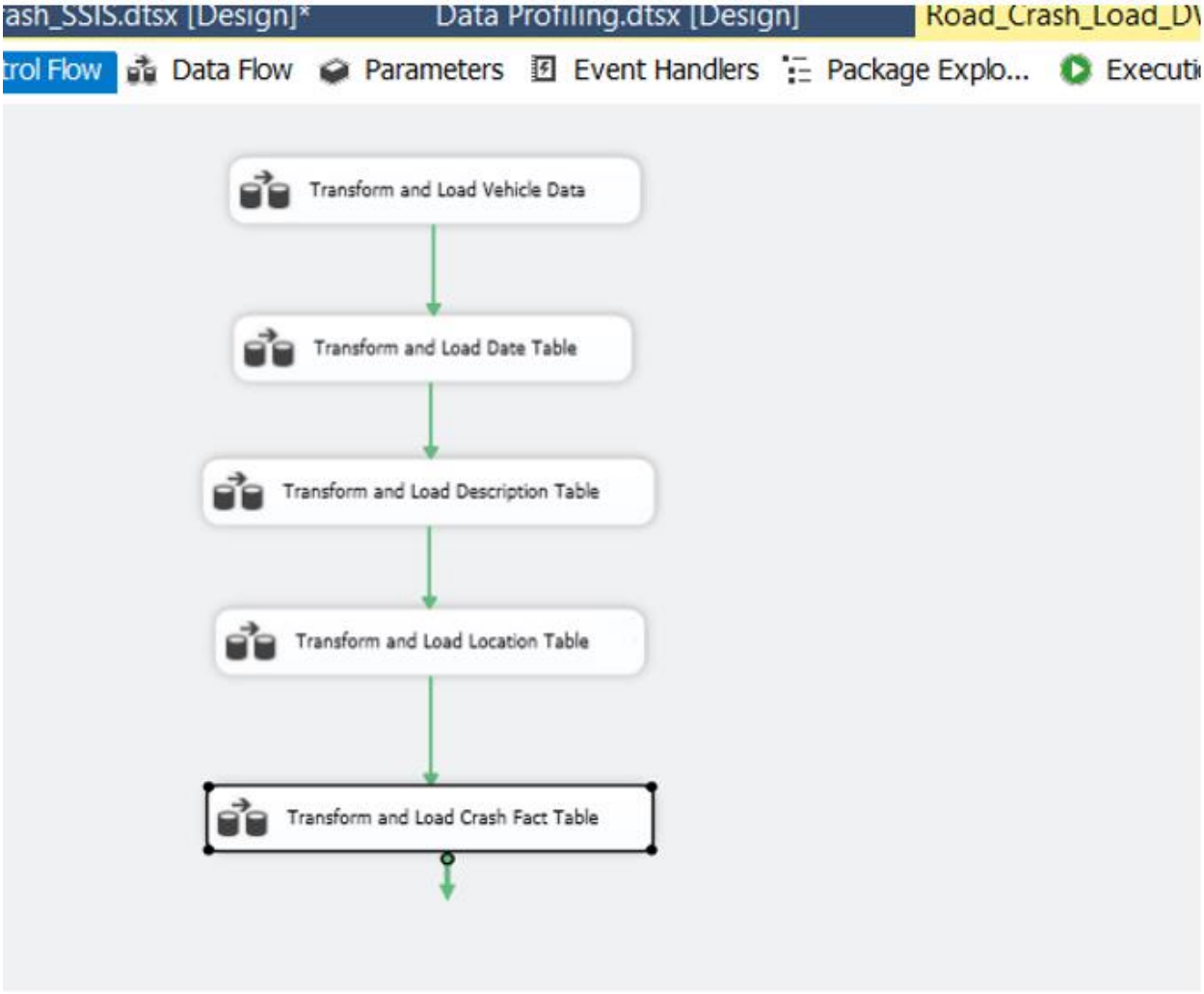

Road_Crash_DW
Execute
Debug
SQLQuery14.sql -...NT2FRU\heesh (56))
ReportServer$MSSQLSERVER_INS2TempDB
Road_Crash_DW
Database Diagrams
Tables
Views
External Resources
Synonyms
Programmability
Stored Procedures
System Stored Procedures
dbo.UpdateDimDate
dbo.UpdateDimDescription
dbo.UpdateDimLocation
dbo.UpdateDimVehicle
Functions
Database Triggers
Assemblies
Types
Rules
Defaults
Plan Guides
Sequences
Service Broker
Storage
Security
Road_Crash_SourceDB
Database Diagrams
Tables
System Tables
FileTables
External Tables
dbo.Crash
dbo.DateTime
dbo.Description
CREATE PROCEDURE [dbo].[UpdateDimVehicle] @vehicles_id NVARCHAR(50)
,@animals INT
,@car_sedan INT
,@car_utility INT
,@car_van INT
,@car_4x4 INT
,@car_station_wagon INT
,@motor_cycle INT
,@truck_small INT
,@truck_large INT
,@bus INT
,@taxi INT
,@bicycle INT
,@scooter INT
,@pedestrian INT
,@inanimate INT
,@train INT
,@tram INT
,@vehicle_other INT
AS
BEGIN
IF NOT EXISTS (
SELECT vehicleSK
FROM dbo.DimVehicle
WHERE vehicleAlternateId = @vehicles_id
)
BEGIN
INSERT INTO dbo.DimVehicle (
[vehicleAlternateId]
,[animals]
,[car_sedan]
,[car_utility]
,[car_van]
,[car_4x4]
,[car_station_wagon]
,[motor_cycle]
,[truck_small]
,[truck_large]
,[bus]
,[taxi]
90 %


Road_Crash_DW
Execute
Debug
SQLQuery14.sql -...NT2FRU\heesh (56))
ReportServer$MSSQLSERVER_INS2TempDB
Road_Crash_DW
Database Diagrams
Tables
Views
External Resources
Synonyms
Programmability
Stored Procedures
System Stored Procedures
dbo.UpdateDimDate
dbo.UpdateDimDescription
dbo.UpdateDimLocation
dbo.UpdateDimVehicle
Functions
Database Triggers
Assemblies
Types
Rules
Defaults
Plan Guides
Sequences
Service Broker
Storage
Security
Road_Crash_SourceDB
Database Diagrams
Tables
System Tables
FileTables
External Tables
dbo.Crash
dbo.DateTime
,@inanimate
,@train
,@tram
,@vehicle_other
,GETDATE()
,GETDATE()
)
END;
IF EXISTS (
SELECT vehicleSK
FROM dbo.DimVehicle
WHERE vehicleAlternateId = @vehicles_id
)
BEGIN
UPDATE dbo.DimVehicle
SET animals = @animals
,car_sedan = @car_sedan
,car_utility = @car_utility
,car_van = @car_van
,car_4x4 = @car_4x4
,car_station_wagon = @car_station_wagon
,motor_cycle = @motor_cycle
,truck_small = @truck_small
,truck_large = @truck_large
,bus = @bus
,taxi = @taxi
,bicycle = @bicycle
,scooter = @scooter
,pedestrian = @pedestrian
,inanimate = @inanimate
,train = @train
,tram = @tram
,vehicle_other = @vehicle_other
,ModifiedDate = GETDATE()
WHERE vehicleAlternateId = @vehicles_id
END;
END;
GO

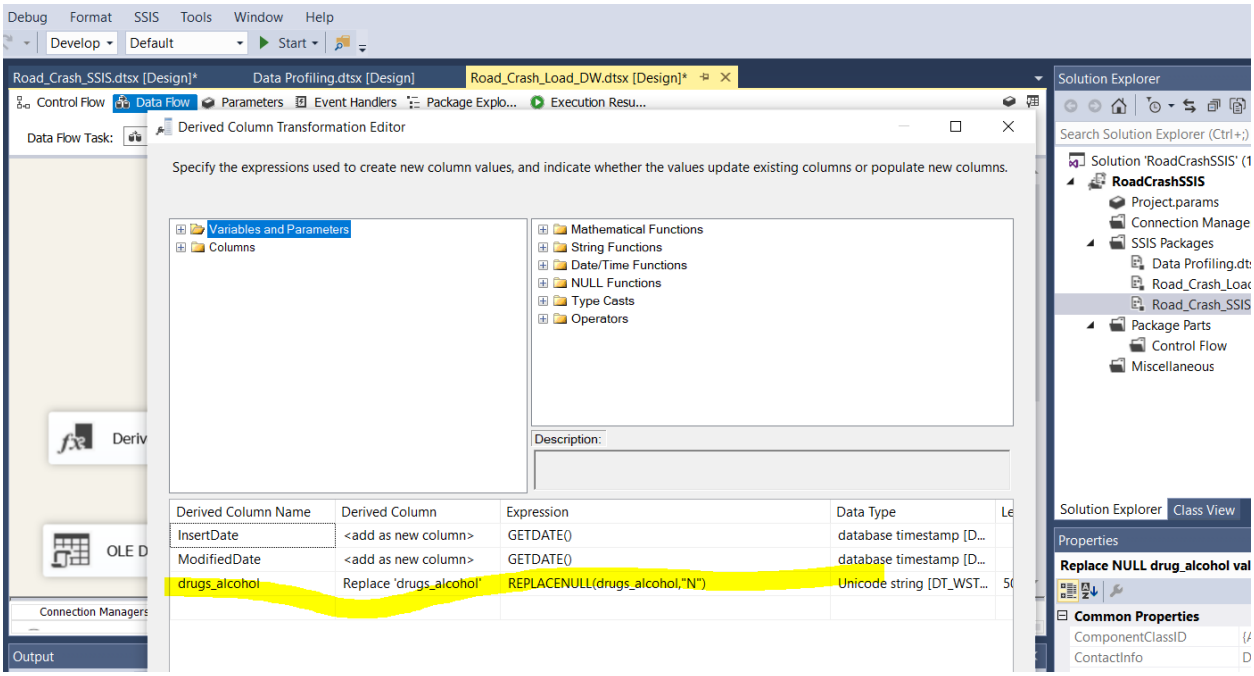
This is how I extracted data from the text file to Staging

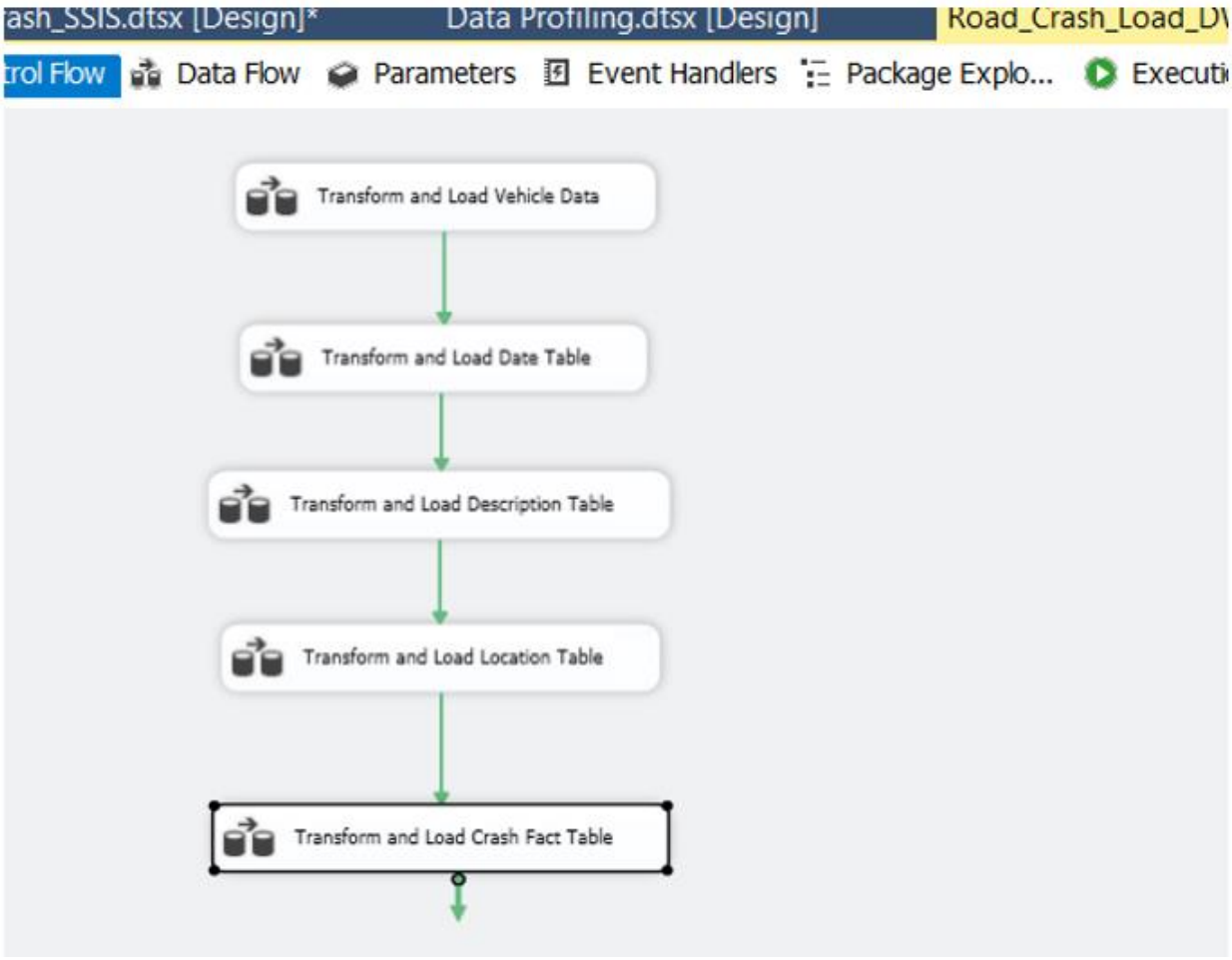

Data Profiling.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Data Flow Task:
Extract Casualties Data to Staging
Flat File Source
OLE DB Destination
Connection Managers
DESKTOP-ENT2FRU\MSSQLSERVER_INS2.Road_C
DESKTOP-ENT2FRU\MSSQLSERVER_INS2.Road_C
Debug
Information: 0x40043007 at Transform and Load Descripti
Information: 0x4004300C at Transform and Load Descripti
Information: 0x40043008 at Transform and Load Descripti
Information: 0x4004300B at Transform and Load Descripti
Information: 0x40043009 at Transform and Load Descripti
SSIS breakpoint 'Break when the container receives the
The program '[26772] DtsDebugHost.exe: DTS' has exited
OLE DB Destination Editor
Configure the properties used to insert data into a relational database using an OLE DB provider.
Connection Manager
Mappings
Error Output
Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder. For fast-load data access, set the table update options.
OLE DB connection manager:
DESKTOP-ENT2FRU\MSSQLSERVER_INS2.Road_Crash_Staging
New...
Data access mode:
Table or view - fast load
Name of the table or the view:
[StgCasualties]
New...
Keep identity
Table lock
Keep nulls
Check constraints
Rows per batch:
Maximum insert commit size:
2147483647
View Existing Data...
OK
Cancel
Help

Extracting data from Staging to Data Warehouse

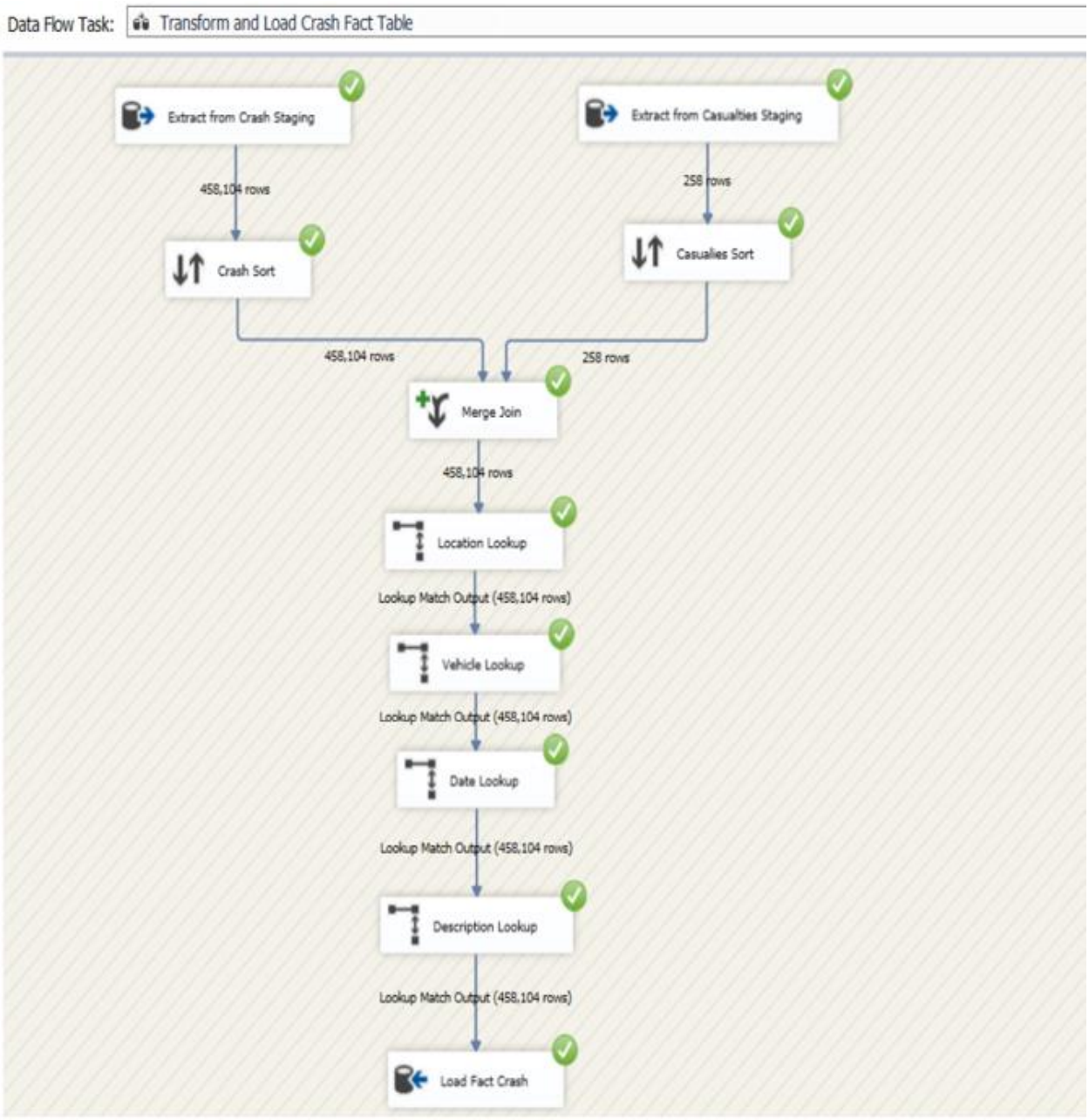


My use of slowly changing dimension on DimDescription. Also I have used derived column in order to replace Null values with N for drugs_alcohol field in DimDescription

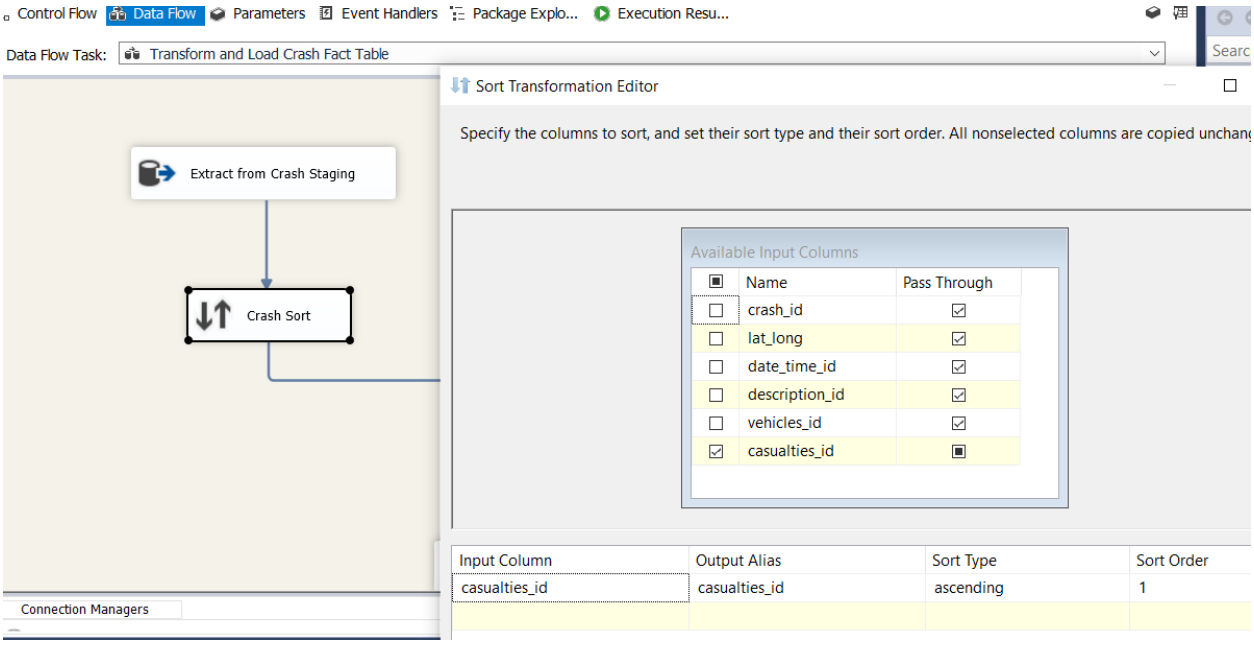


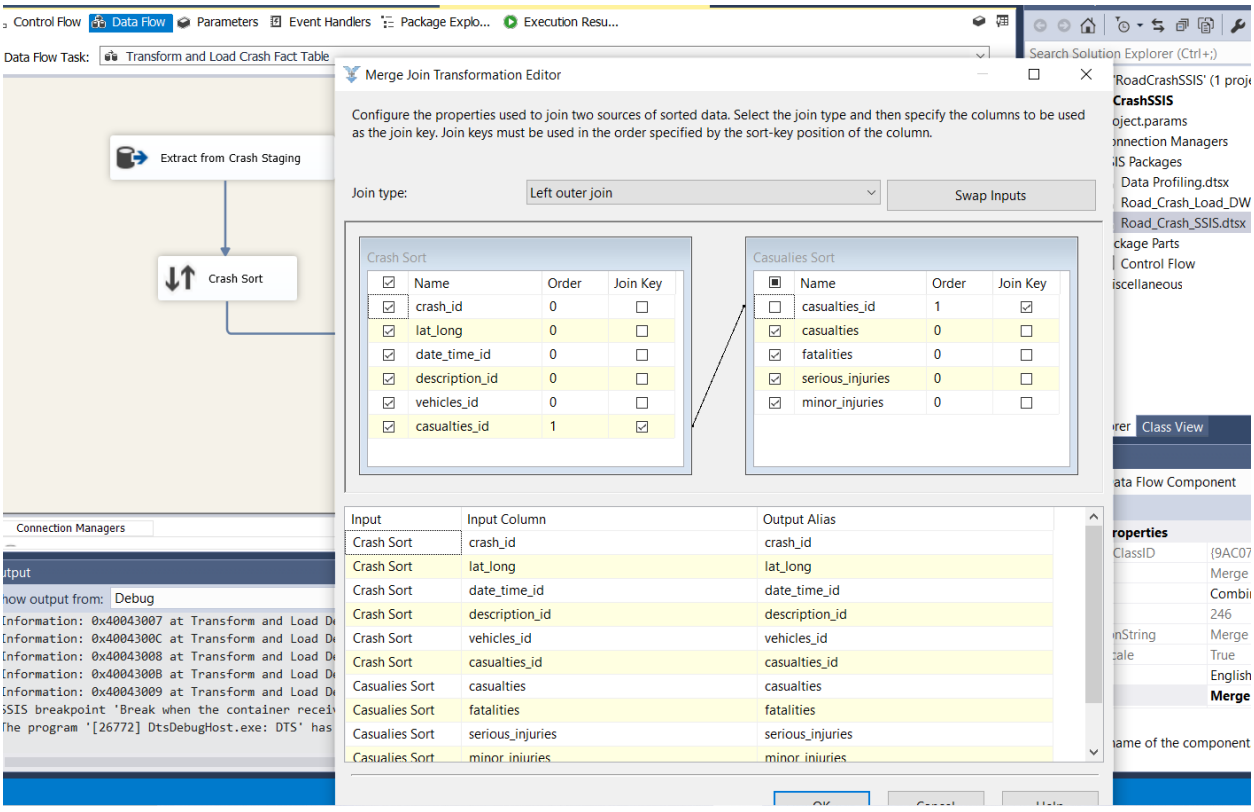


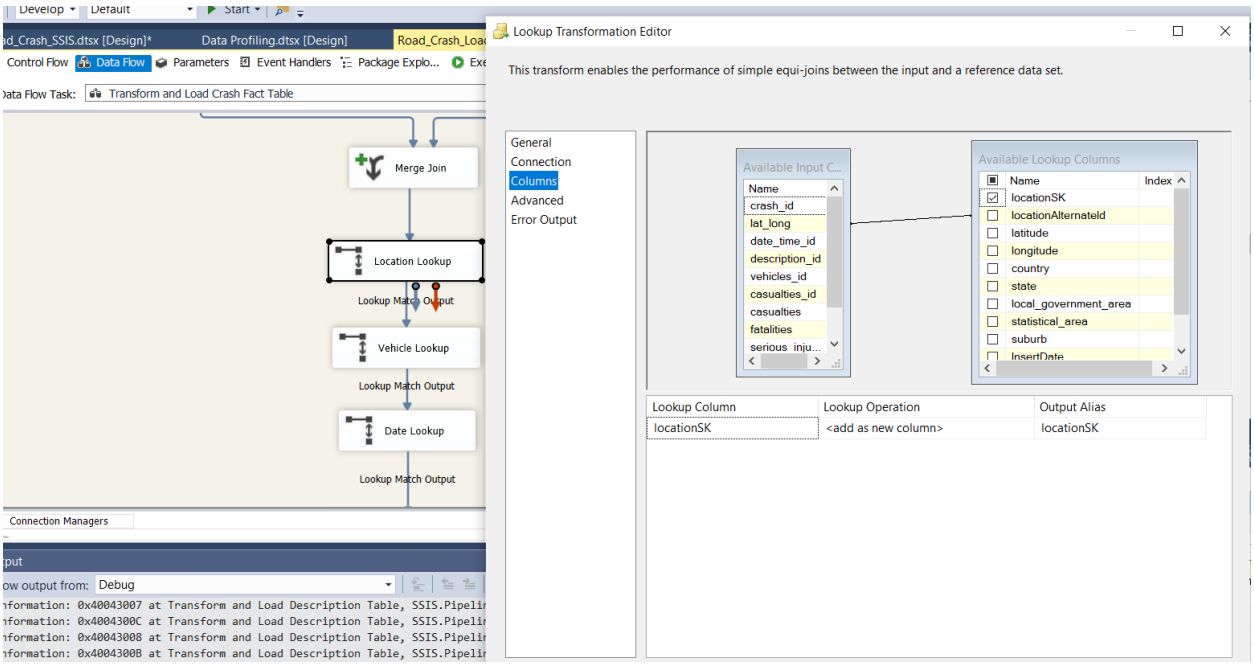
Below I have merged the extracted crash table with casualties table with crash table sorting them with casualties_id and then I have used for lookups in order to load the transformed data into data warehouse with the use of surrogate keys



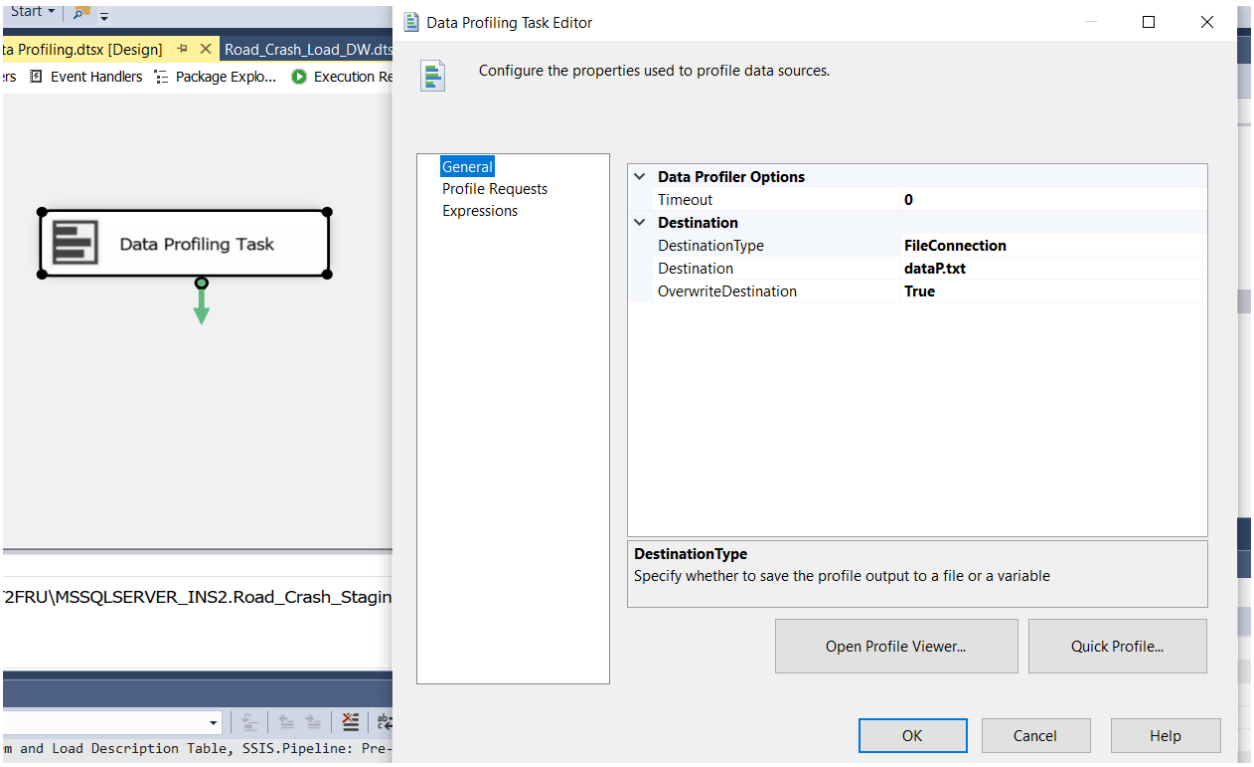
Below is Fact Crash Table in Data Warehouse

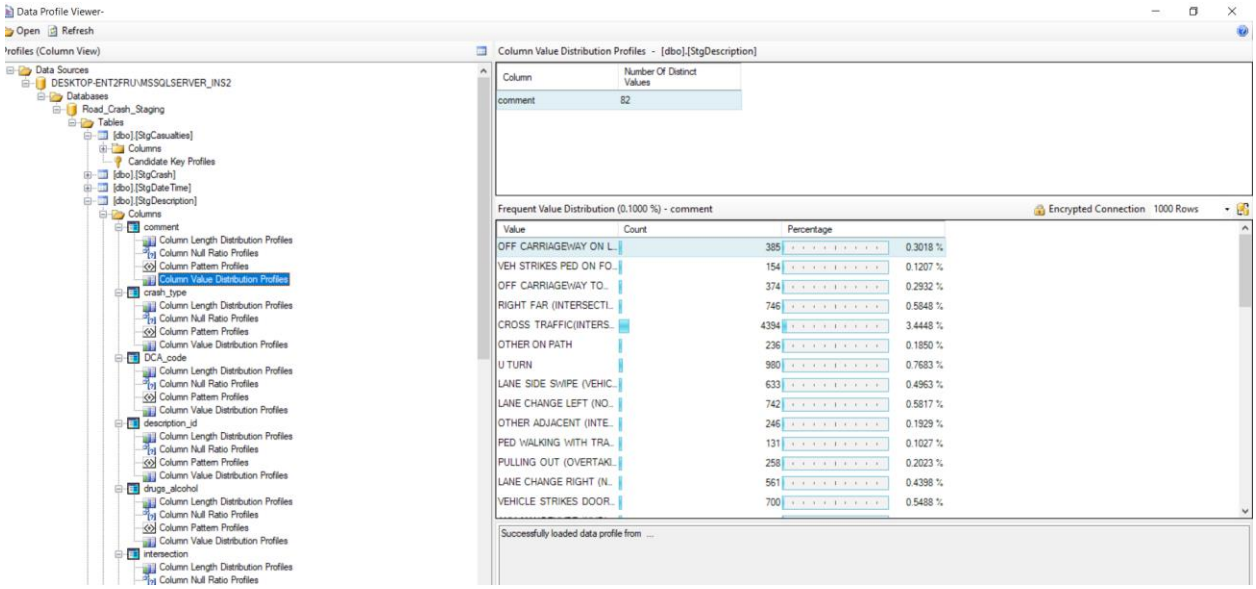

Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Execution Resu...
Data Flow Task:
Transform and Load Crash Fact Table
Extract from Crash Staging
Crash Sort
Connection Managers
Sort Transformation Editor
Specify the columns to sort, and set their sort type and their sort order. All nonselected columns are copied unchang
Available Input Columns
Name
Pass Through
crash_id
lat_long
date_time_id
description_id
vehicles_id
casualties_id
Input Column
Output Alias
Sort Type
Sort Order
casualties_id
casualties_id
ascending
1


Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Execution Resu...
Data Flow Task:
Transform and Load Crash Fact Table
Extract from Crash Staging
Crash Sort
Connection Managers
Merge Join Transformation Editor
Configure the properties used to join two sources of sorted data. Select the join type and then specify the columns to be used as the join key. Join keys must be used in the order specified by the sort-key position of the column.
Join type:
Left outer join
Swap Inputs
Crash Sort
Name
Order
Join Key
crash_id
lat_long
date_time_id
description_id
vehicles_id
casualties_id
Casualties Sort
casualties_id
casualties
fatalities
serious_injuries
minor_injuries
Input
Input Column
Output Alias
Search Solution Explorer (Ctrl+;)
Class View
Debug



Data Profiling




Data Profile Viewer-
Open
Refresh
Profiles (Column View)
Data Sources
DESKTOP-ENT2FRU\MSSQLSERVER_INS2
Databases
Road_Crash_Staging
Tables
[dbo].[StgCasualties]
Columns
Candidate Key Profiles
[dbo].[StgCrash]
[dbo].[StgDateTime]
[dbo].[StgDescription]
Columns
comment
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Value Distribution Profiles
crash_type
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Value Distribution Profiles
DCA_code
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Value Distribution Profiles
description_id
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Value Distribution Profiles
drugs_alcohol
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Value Distribution Profiles
intersection
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Value Distribution Profiles - [dbo].[StgDescription]
Column
Number Of Distinct Values
comment
82
Frequent Value Distribution (0.1000 %) - comment
Encrypted Connection
1000 Rows
Value
Count
Percentage
OFF CARRIAGEWAY ON L.. 385 0.3018 %
VEH STRIKES PED ON FO.. 154 0.1207 %
OFF CARRIAGEWAY TO.. 374 0.2932 %
RIGHT FAR (INTERSECTI.. 746 0.5848 %
CROSS TRAFFIC(INTERS.. 4394 3.4448 %
OTHER ON PATH 236 0.1850 %
U TURN 980 0.7683 %
LANE SIDE SWIPE (VEHIC.. 633 0.4963 %
LANE CHANGE LEFT (NO.. 742 0.5817 %
OTHER ADJACENT (INTE.. 246 0.1929 %
PED WALKING WITH TRA.. 131 0.1027 %
PULLING OUT (OVERTAKI.. 258 0.2023 %
LANE CHANGE RIGHT (N.. 561 0.4398 %
VEHICLE STRIKES DOOR.. 700 0.5488 %
Successfully loaded data profile from ...