Data Warehousing & Business Intelligence

New York City Taxi Trips

Part 1

Developed by:

D.R.A KUMARAGE IT16058156

Submitted to:

MR. Sheron Dinushka

Table of Contents

1. Data set selection

Background

This scenario is mainly based the new york city taxi trips database. In this particular scenario Customer, Vehicle and Driver details are not available, Thereby I use data from The Chicago Taxi details for connect to my main database.

Uber,Pickme like taxi services scenario are examples for the above scenario

Customer can go many number of trips , each vehicle Assigned a one driver. Trip detail data base has all trip details

Content

The data was downloaded By exact link:

Chicago Taxi Trips :

https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew

new-york-city-taxi-trips

https://www.kaggle.com/kentonnlp/2014-new-york-city-taxi-trips

Driver Details

https://data.cityofchicago.org/Community-Economic-Development/Public-Chauffeurs/97wa-y6ff

1. 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)

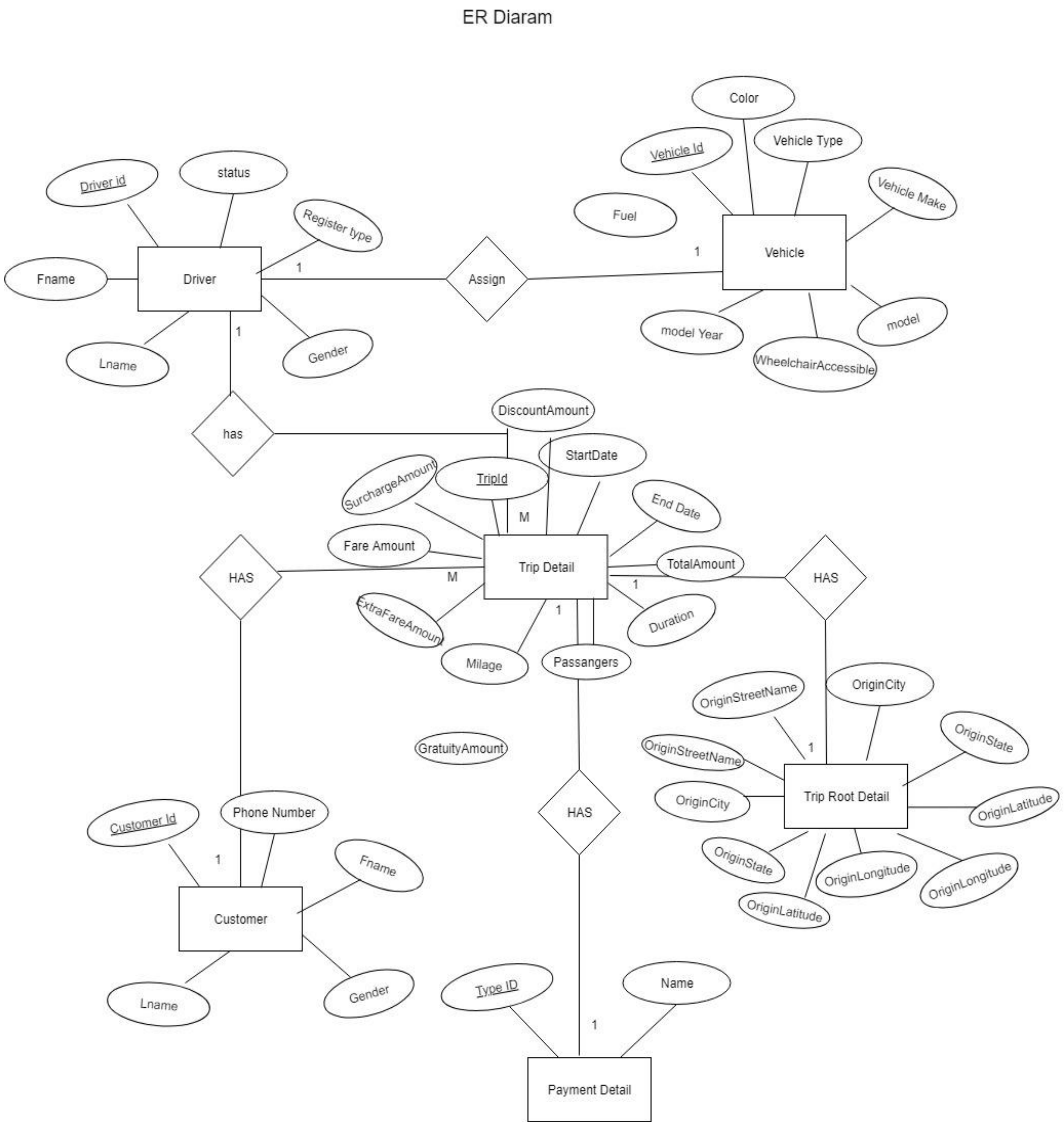
2. Text file (.txt)

3. Excel file (.xlsx)

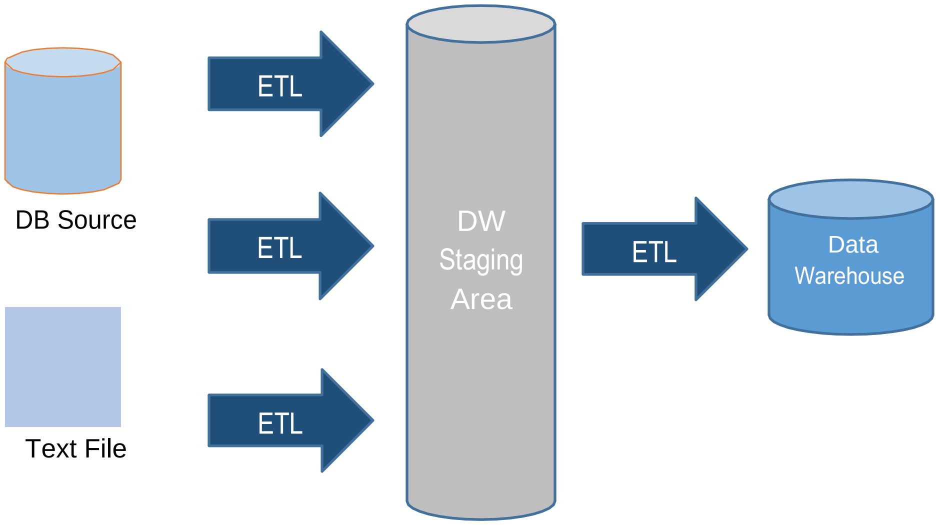
Text file: Customer Address, Driver Address

Excel file: Customer , Driver , Trip Detail , Trip Root Detail , Vehicle .

ER Diagram


ER Diaram
Driver id
status
Register type
Fname
Driver
Lname
Gender
1
1
Assign
Fuel
Color
Vehicle Id
Vehicle Type
Vehicle Make
1
Vehicle
model Year
WheelchairAccessible
model
has
DiscountAmount
StartDate
SurchargeAmount
TripId
End Date
M
Fare Amount
Trip Detail
TotalAmount
HAS
M
1
HAS
ExtraFareAmount
1
Duration
Milage
Passangers
OriginStreetName
OriginCity
1
OriginStreetName
OriginState
GratuityAmount
OriginCity
Trip Root Detail
OriginLatitude
HAS
OriginState
OriginLongitude
OriginLongitude
OriginLatitude
Customer Id
Phone Number
1
Fname
Customer
Lname
Gender
Type ID
Name
1
Payment Detail

3. Solution Architecture



- Customer Address Staging
- Customer Staging
- Driver Staging
- Driver Address Staging
- Vehicle Staging
- Payment Type Staging
- Trip Detail Staging
- Trip Root Detail 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

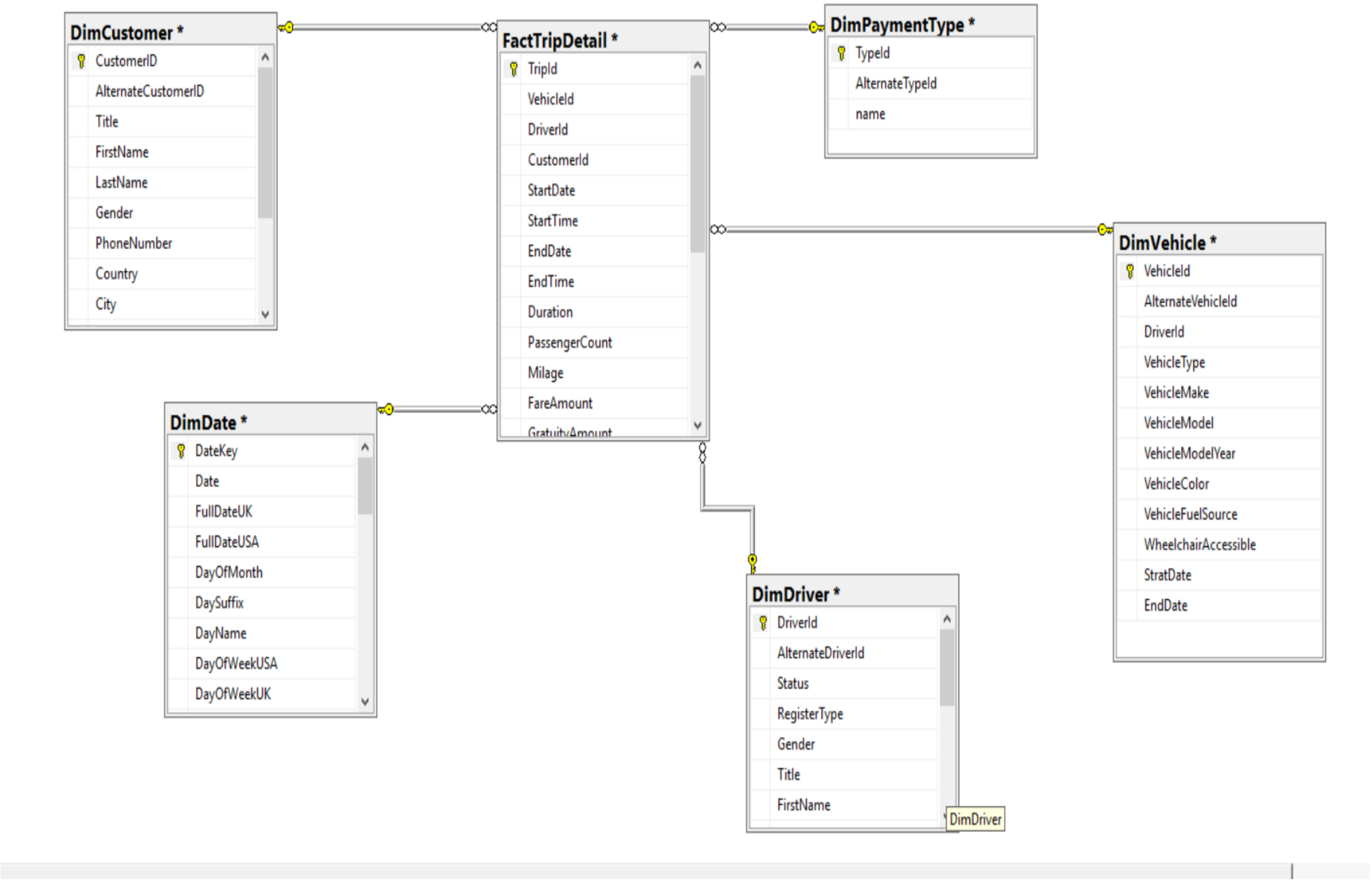
- **Data Warehouse**

EDW vs Data Mart

Dimensional Modeling - Facts & Dimension

Many Schemas -In here used Star schema

4. Data warehouse design & development



DimCustomer *
- CustomerID
- AlternateCustomerID
- Title
- FirstName
- LastName
- Gender
- PhoneNumber
- Country
- City

FactTripDetail *
- TripId
- VehicleId
- DriverId
- CustomerId
- StartDate
- StartTime
- EndDate
- EndTime
- Duration
- PassengerCount
- Milage
- FareAmount
- GratuityAmount

DimPaymentType *
- TypeId
- AlternateTypeId
- name

DimVehicle *
- VehicleId
- AlternateVehicleId
- DriverId
- VehicleType
- VehicleMake
- VehicleModel
- VehicleModelYear
- VehicleColor
- VehicleFuelSource
- WheelchairAccessible
- StratDate
- EndDate

DimDate *
- DateKey
- Date
- FullDateUK
- FullDateUSA
- DayOfMonth
- DaySuffix
- DayName
- DayOfWeekUSA
- DayOfWeekUK

DimDriver *
- DriverId
- AlternateDriverId
- Status
- RegisterType
- Gender
- Title
- FirstName

5. ETL development

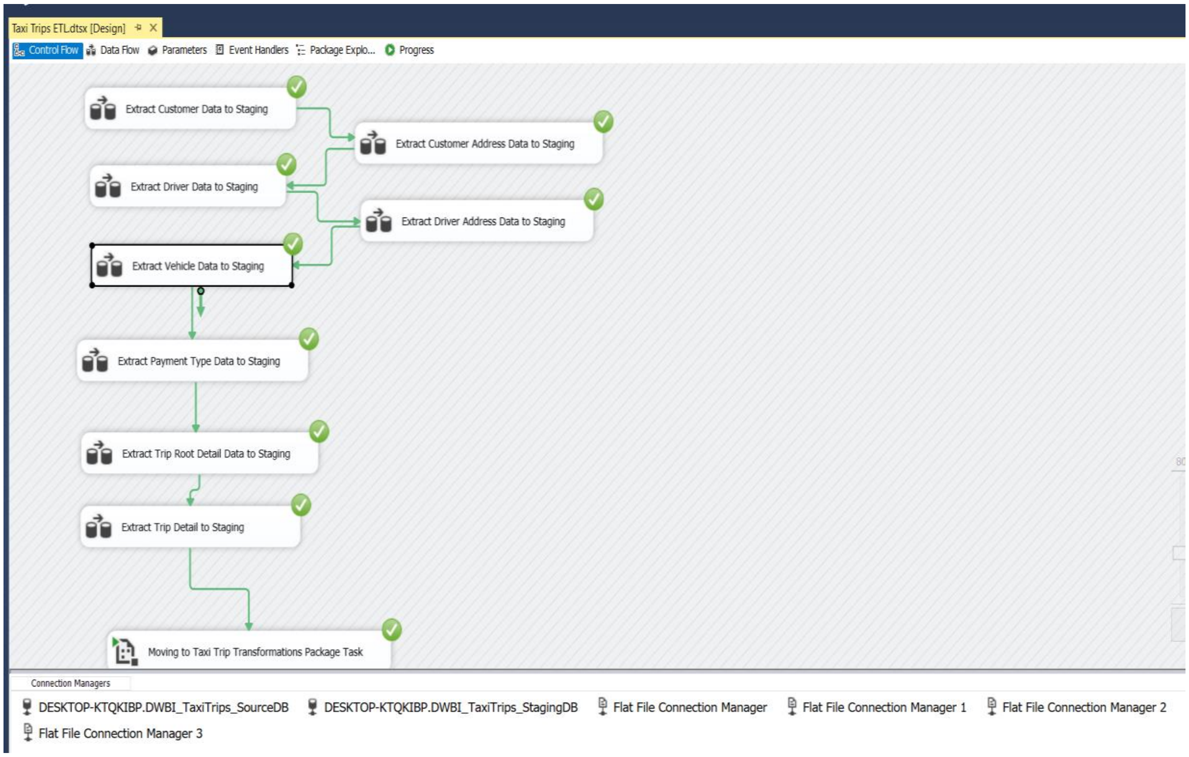
ETL (Extract-Transform-Load)

ETL comes from Data Warehousing and stands for Extract-Transform-Load. ETL covers a process of how the data are loaded from the source system to the data warehouse. Currently, the ETL encompasses a cleaning step as a separate step. The sequence is then Extract-Clean-Transform Load. Let us briefly describe each step of the ETL process.

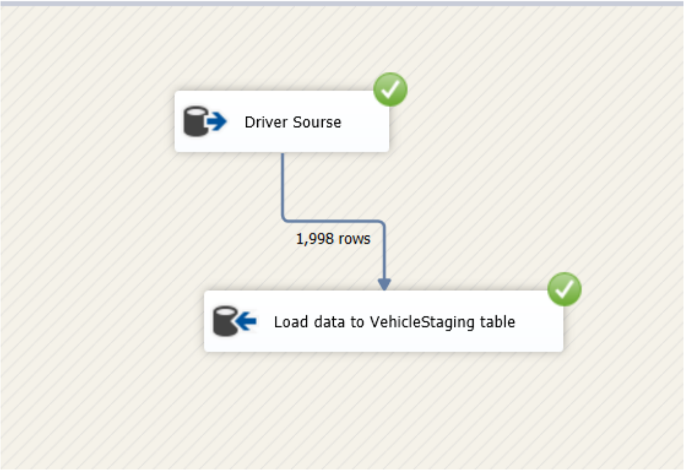
- **Extract**

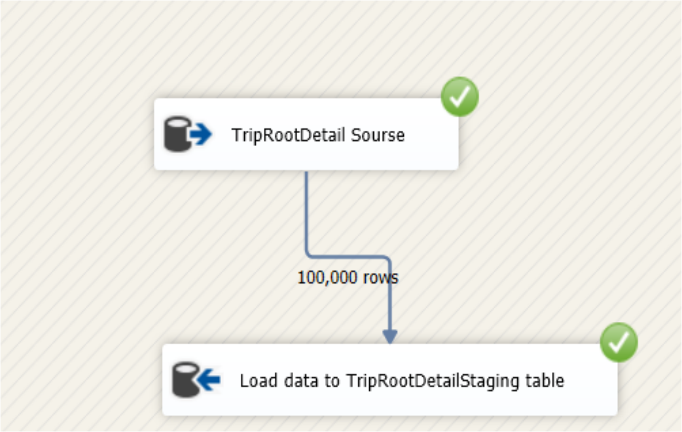
The main objective of the extract step is to retrieve all the required data from the source system with as little resources as possible.

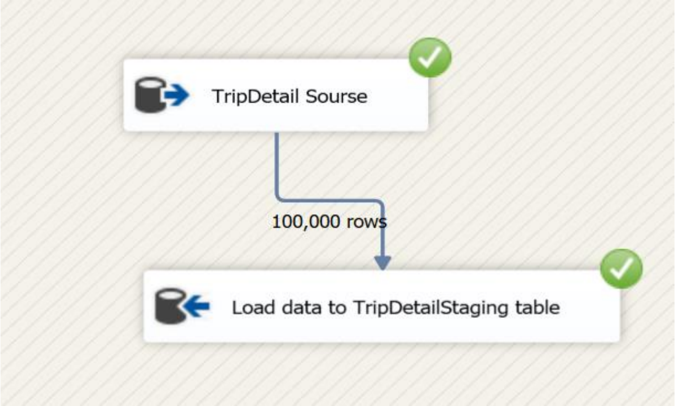
Data Extraction from Sources to Staging Tables

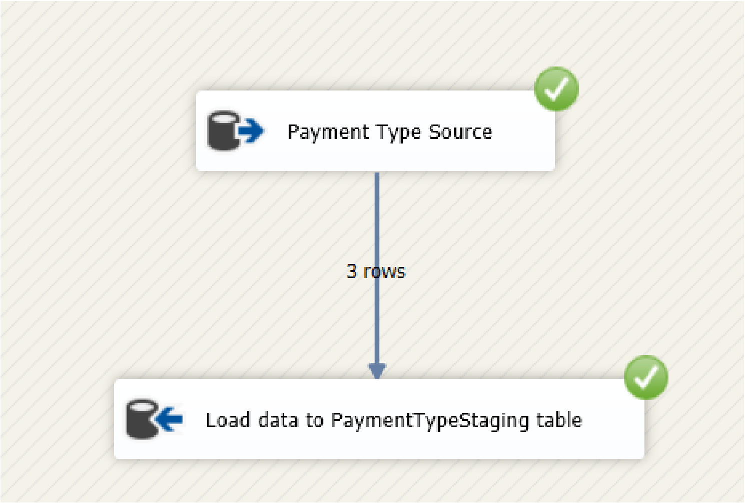


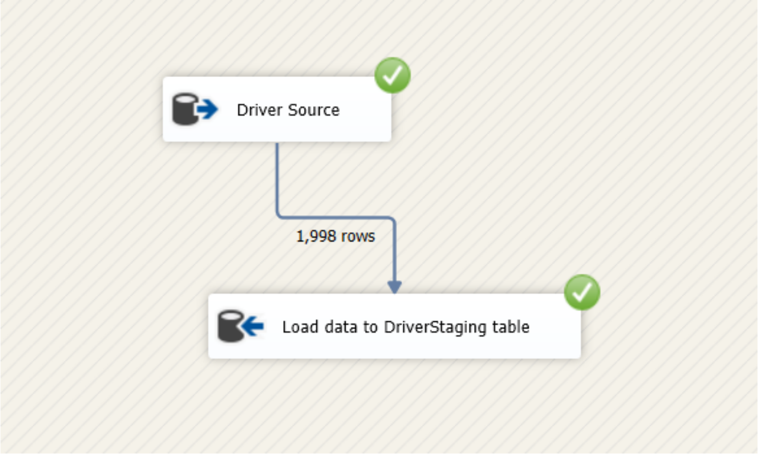
Taxi Trips ETL.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Extract Customer Data to Staging
Extract Customer Address Data to Staging
Extract Driver Data to Staging
Extract Driver Address Data to Staging
Extract Vehicle Data to Staging
Extract Payment Type Data to Staging
Extract Trip Root Detail Data to Staging
Extract Trip Detail to Staging
Moving to Taxi Trip Transformations Package Task
Connection Managers
DESKTOP-KTQKIBP.DWBI_TaxiTrips_SourceDB
DESKTOP-KTQKIBP.DWBI_TaxiTrips_StagingDB
Flat File Connection Manager
Flat File Connection Manager 1
Flat File Connection Manager 2
Flat File Connection Manager 3

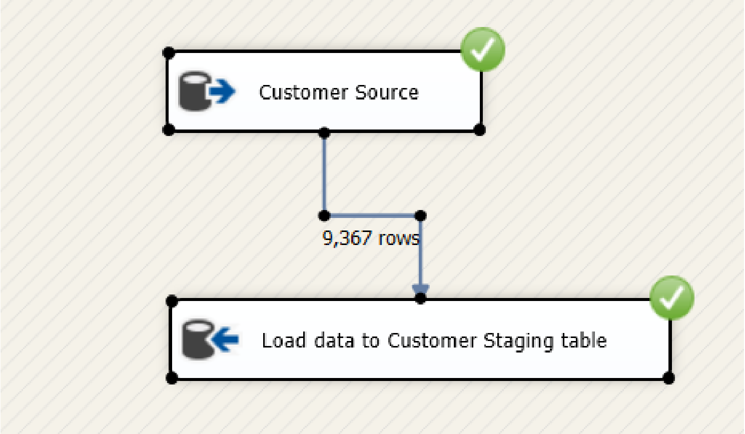

Driver Sourse
1,998 rows
Load data to VehicleStaging table

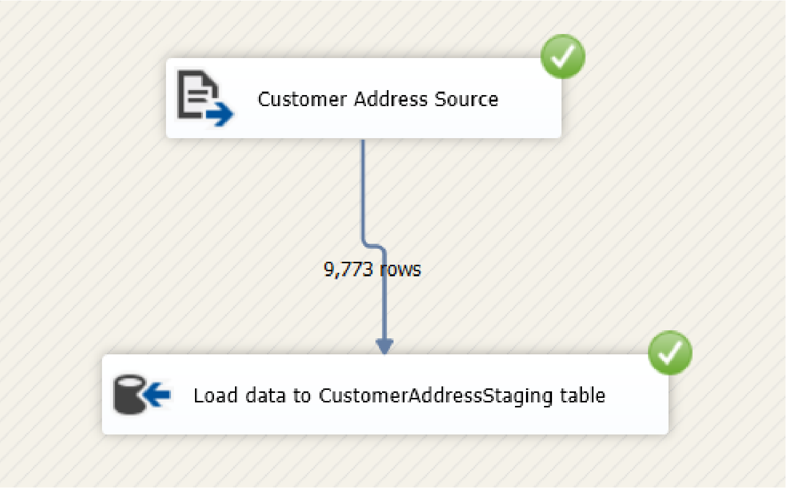

TripRootDetail Sourse
100,000 rows
Load data to TripRootDetailStaging table


TripDetail Sourse
100,000 rows
Load data to TripDetailStaging table


Payment Type Source
3 rows
Load data to PaymentTypeStaging table

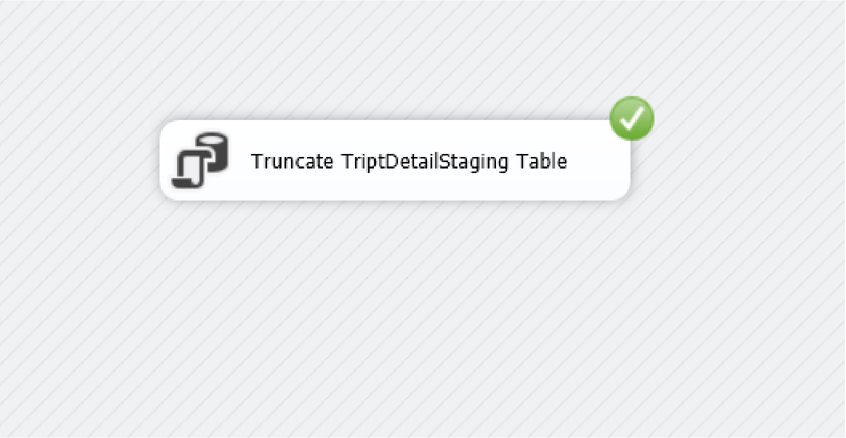

Driver Source
1,998 rows
Load data to DriverStaging table


Customer Source
9,367 rows
Load data to Customer Staging table

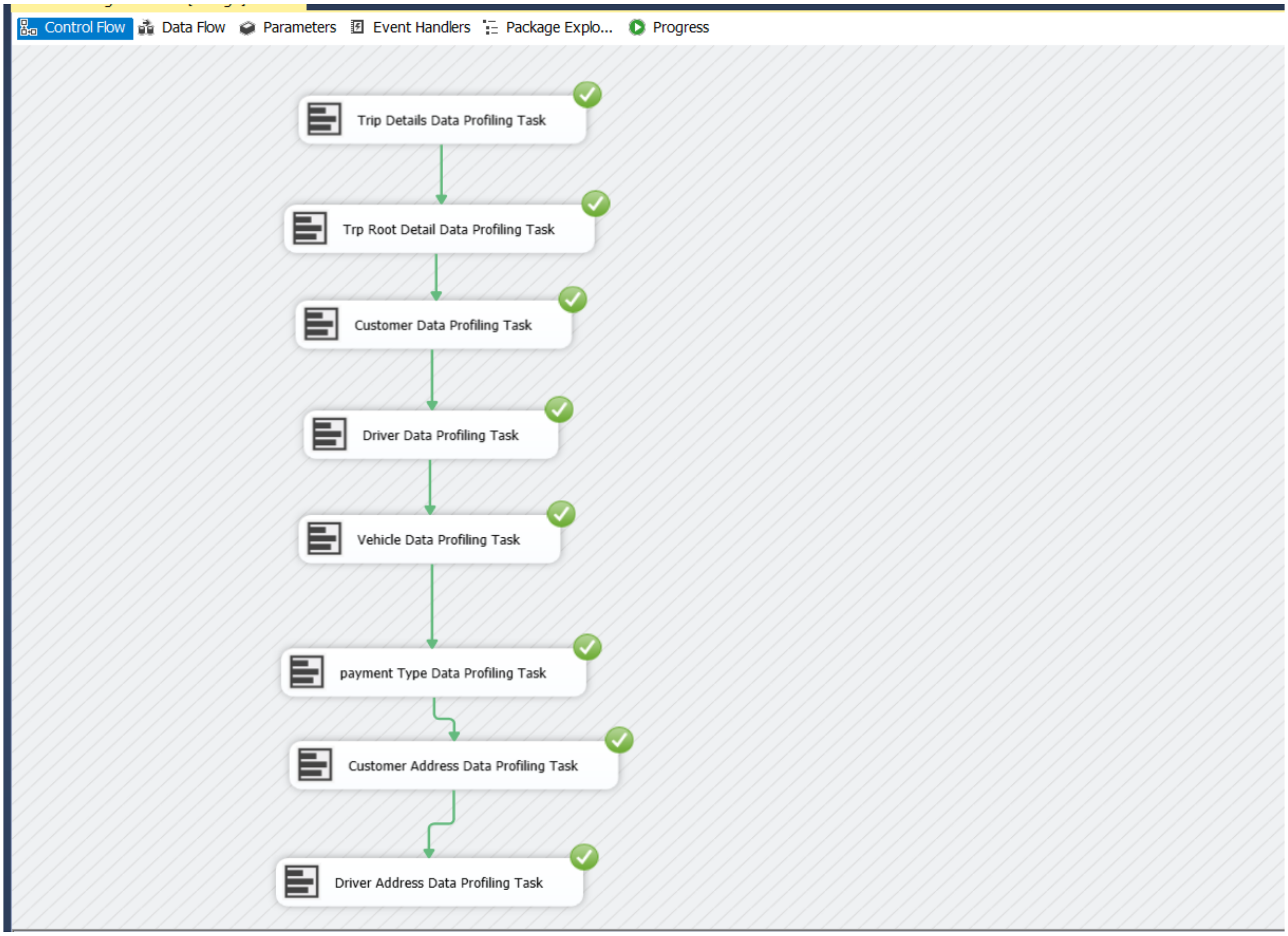


Event Handler In Staging ETL

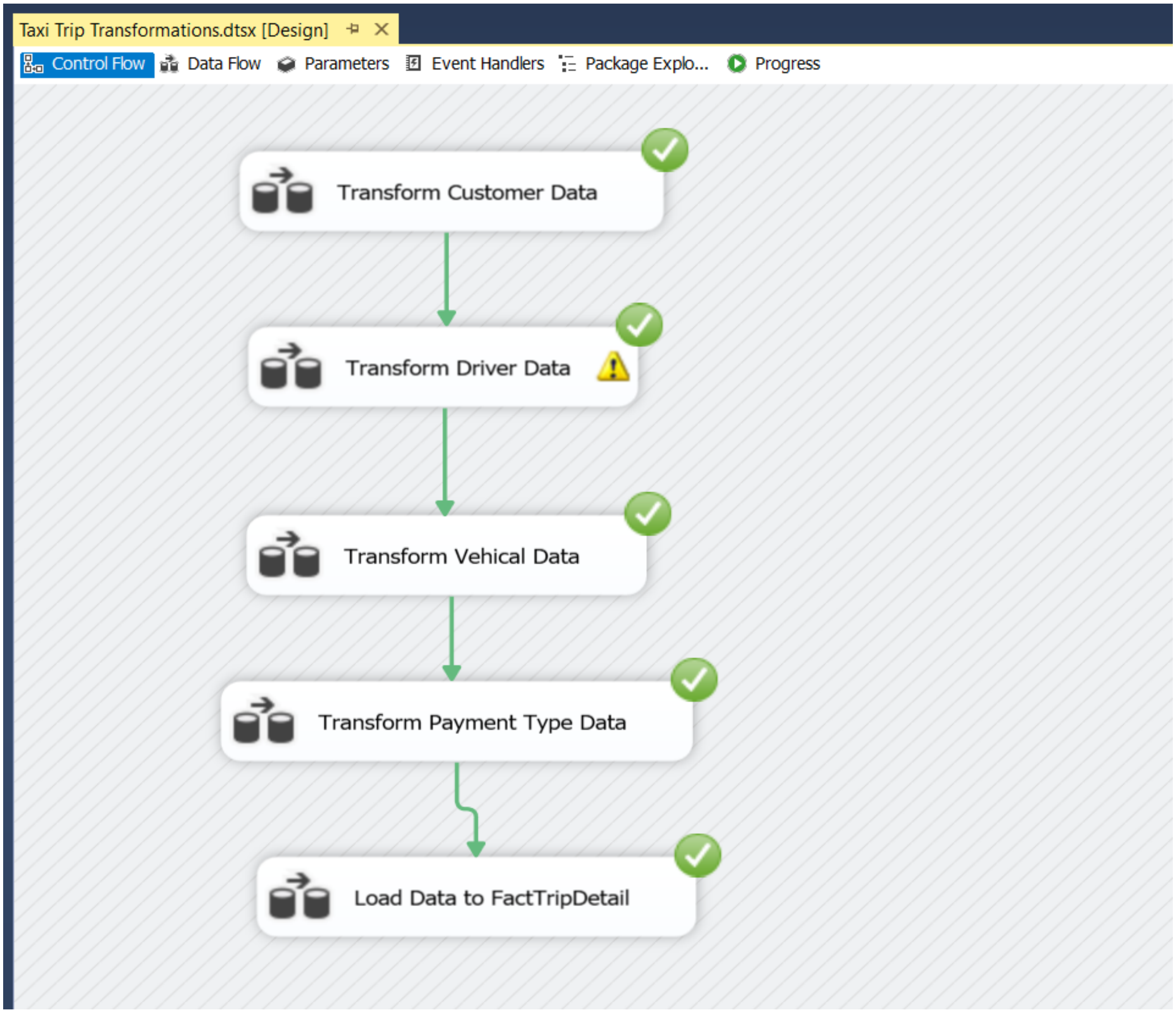
EX:



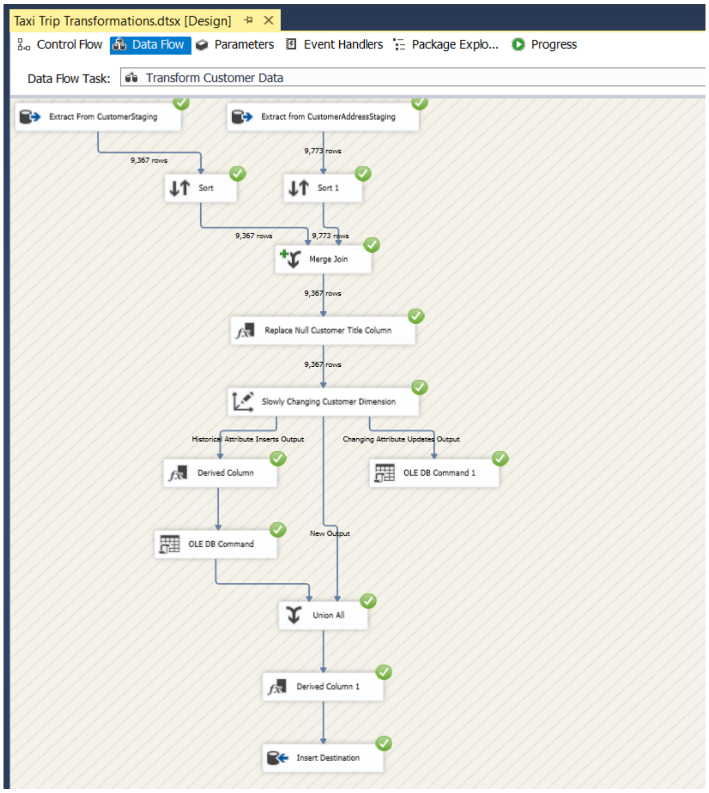
Data Profiling


Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Trip Details Data Profiling Task
Trp Root Detail Data Profiling Task
Customer Data Profiling Task
Driver Data Profiling Task
Vehicle Data Profiling Task
payment Type Data Profiling Task
Customer Address Data Profiling Task
Driver Address Data Profiling Task

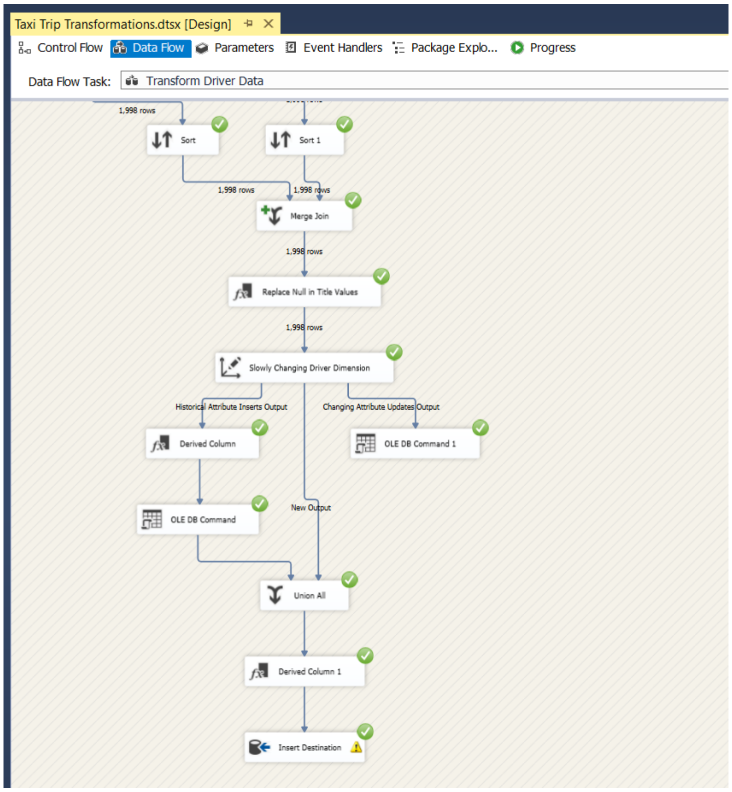
Data Transformation Control Flow



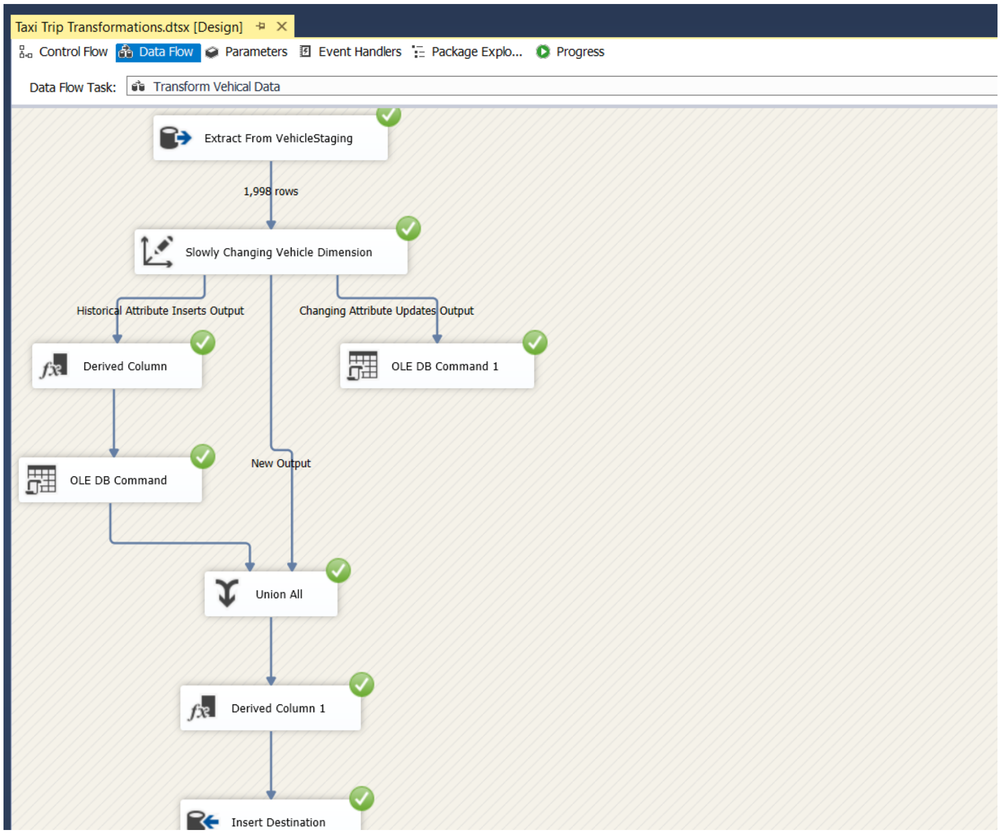
Data Flow of Loading Dimension and Loading Fact Table



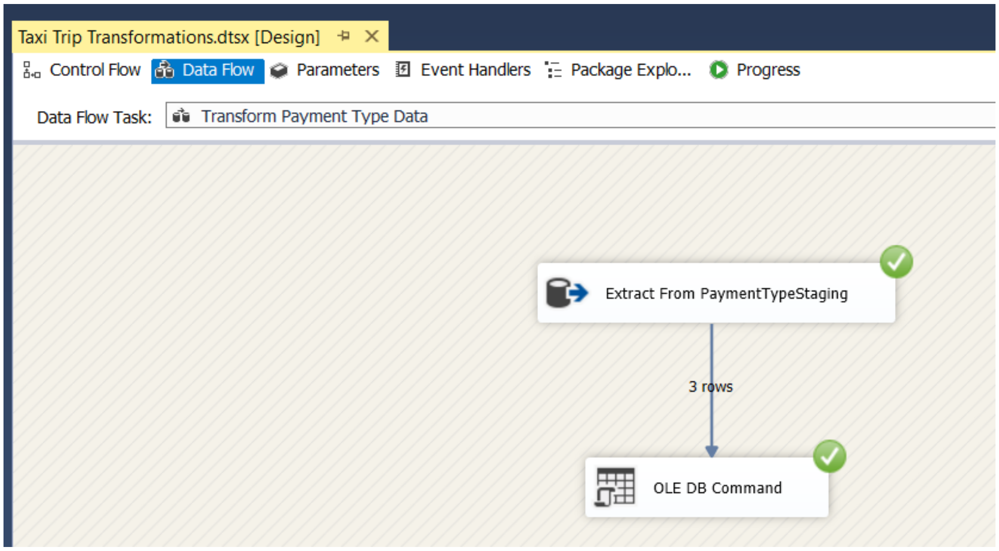
Data Flow Task: Transform Customer Data



Data Flow Task: Transform Driver Data

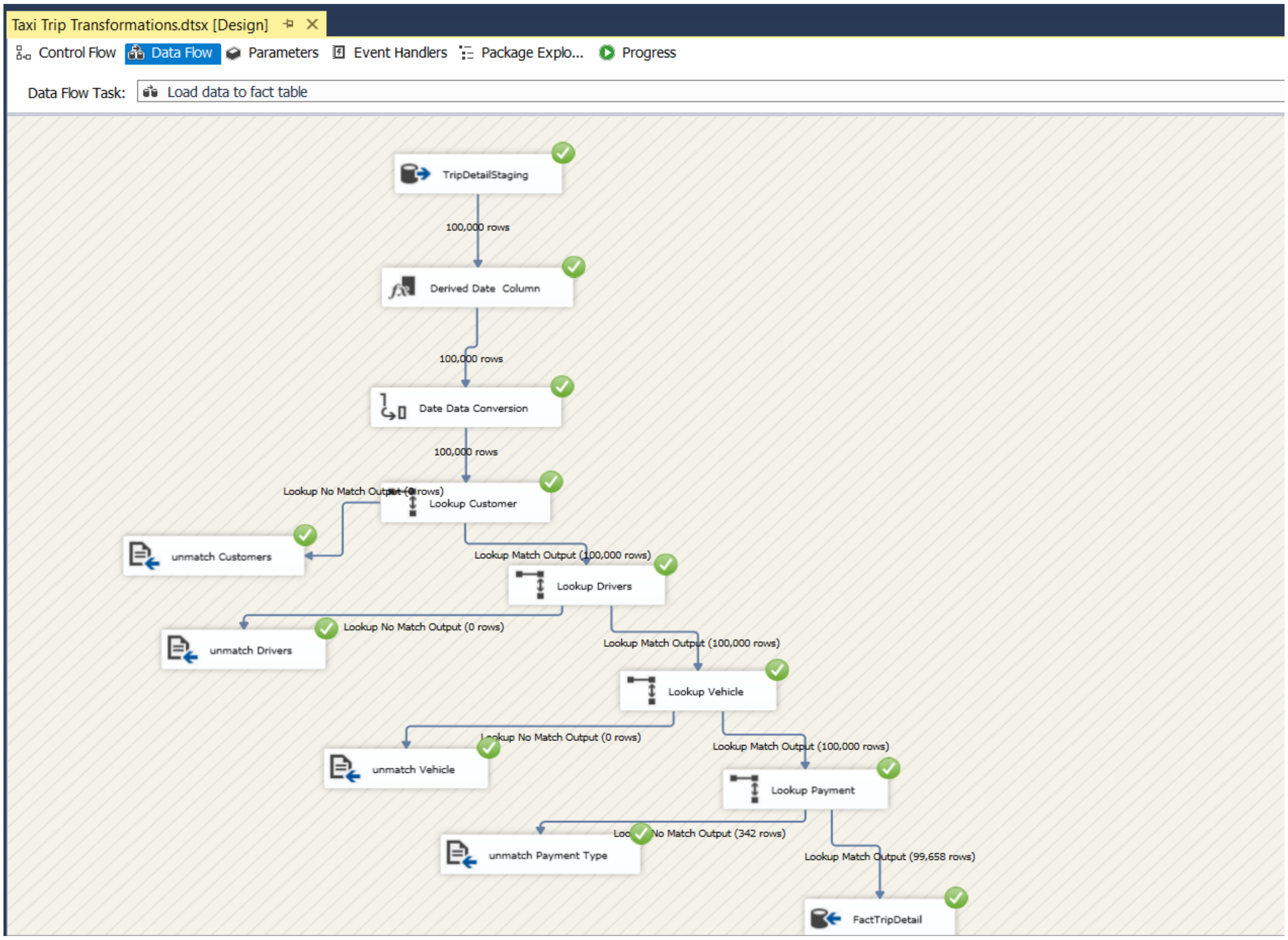


Data Flow Task: Transform Vehical Data



Data Flow Task: Transform Payment Type Data

Load Fact Table :



Taxi Trip Transformations.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Data Flow Task:
Load data to fact table
TripDetailStaging
100,000 rows
Derived Date Column
100,000 rows
Date Data Conversion
100,000 rows
Lookup No Match Output (0 rows)
Lookup Customer
unmatch Customers
Lookup Match Output (100,000 rows)
Lookup Drivers
Lookup No Match Output (0 rows)
unmatch Drivers
Lookup Match Output (100,000 rows)
Lookup Vehicle
Lookup No Match Output (0 rows)
unmatch Vehicle
Lookup Match Output (100,000 rows)
Lookup Payment
Lookup No Match Output (342 rows)
unmatch Payment Type
Lookup Match Output (99,658 rows)
FactTripDetail