IMPACT OF WEATHER & COVID ON ACCIDENTS

Project Proposal

GROUP 09
Hemanth Reddy Musuku
Raja Muthu
Sowmya Shree Nagaraju
Naga Santhosh Kartheek Karnati

Data

We have 3 data sources:

Accidents Data Source: The accidents data consists of traffic accident events in the US from 2019 to June 2020. It consists of attributes like start time, end time, the type of accident, the severity of the accident and some information about the location such a city, state, county, zip code, etc.

Link: https://osu.app.box.com/v/traffic-events-june20

Weather Data Source : Similar to the accidents data, the weather data contains information about weather events in the USA from 2019 to June 2020 about the type, severity, the timestamp and location information, not unlike the Accidents data.

Link : https://osu.app.box.com/v/weather-events-june20

COVID-19 Data Source: In this dataset we have the COVID-19 case details of each county in the US from Jan 2020 to till date. This dataset has details about number of cases, cumulative case per thousand and million, health index, economic support index, how many hospitalized, how many deaths, restricted on gathering, schools closed and many. We are planning to join this dataset along with accidents dataset based on the county and date to check the impact of covid on traffic and accidents.

Link:https://github.com/google-research/open-covid-19-data/tree/master/data/exports

Joining of the datasets:

We hope to merge the Accidents data to the Weather data, by date and location, to understand the relationship between them. Also, we want to investigate the relationship between Covid Cases in a region and the traffic events, whether any important relationship exists between the two. Again, we would join by date and location. It has also been said that the lockdown has improved the environmental conditions – As such, we would join the covid data to the accidents dataset to understand these effects.

Transformation:

There are a lot of unnecessary columns in the 3 data sources, along with apparent errors and Null values.

Weather Data: We will remove the columns Time Zone, Latitude, Longitude and Airport Code

Traffic Data: For the traffic events data, we are going to remove the columns Time Zone, Latitude, Longitude, Airport Code, TMC

Covid Data: For this dataset we have columns with more than 60% null values which will not be necessary for our analysis and these columns will be removed. We will be merging this dataset with another location details dataset to get the county details. We need to make sure only those counties which has the traffic details will be considered for our analysis along with date matching.

Also, we need to ensure the date format is consistent in the 3 data sources, before we merge all the data sources into one. For this, we plan to split the now date-time column into 2 parts – date and time.

The reason for doing this is because there is no information about time in the covid data, since we only have daily data. With this, we can ensure more consistent data throughout the Data Warehouse.

Dimensions:

The dimensions of the data would include primarily information about the date and time, location, weather event and its severity, traffic event data, along with covid information including cases, deaths, demographic information by county to understand the relationships perfectly.

High Level Data Flow:

We plan to clean the 3 data sources separately, that is, remove unnecessary columns and format it properly, then merge them into one data source with the appropriate columns. Next, the data will be loaded into a staging area, and concurrently an archive for the merged data source will also be created. Then, after going through the data in the staging area, we would perform the required transformations and then load it into the master table. After that, once we have the data warehouse sorted, for now, we plan to have at least 2 data marts- Pre-Covid and Post-Covid data, which can then be used for further analysis.

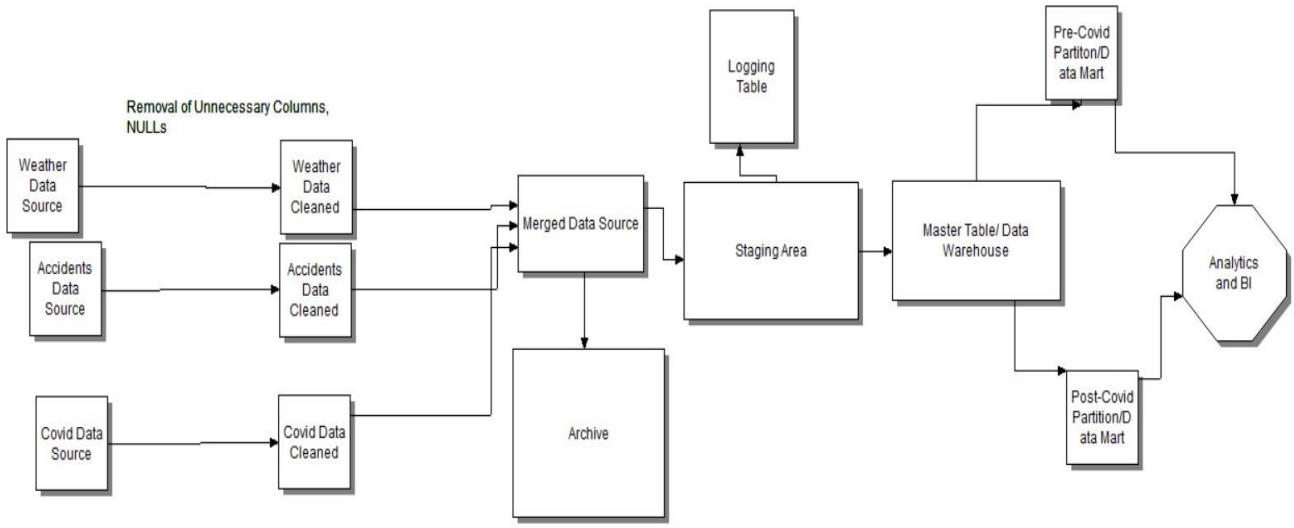


Fig. 1 High level data flow

Error Handling:

Weather Data

EventID:

This will be the Primary Key and no duplicate values will be allowed.

DateTime:

We plan to split the UTC datetime column into date and time. With this, we will have a consistent data type for the columns in data pre and post covid, as the Post Covid data is mostly daily data.

Also, we will implement an outlier handler that will not allow entries with dates before and after certain range, as per the requirement.

Severity/Type:

We are going to impose some constraints on these columns to ensure only valid entries (For Severity – moderate, light, heavy and for Type – Snow, Rain, fog, etc.)

Location Information:

For State, County, City and ZipCode, we will implement constraints to ensure everything is in a consistent format and take care of any data entry errors.

Traffic Data

EventID:

Similar to the traffic data, this will be the Primary Key and no duplicate values will be allowed.

DateTime:

We plan to split the UTC datetime column into date and time. With this, we will have a consistent data type for the columns in data pre and post covid, as the Post Covid data is mostly daily data.

Also, we will implement an outlier handler that will not allow entries with dates before and after certain values, as per the requirement.

Severity/Type:

We are going to impose some constraints on these columns to ensure only valid entries (For Severity – fast, slow moderate and for Type –Congestion, Broken-Vehicle, etc.)

Location Data:

For State, County, City and ZipCode, we will implement constraints to ensure everything is in a consistent format and take care of any data entry errors.

Covid Data

Date Time and County

These two columns will be our combined primary key. And no redundant data with the same combination allowed.

Open covid region code

This column has code of every county with state abbreviation and needs to be split into two columns. One with state and the other with code.

Test units

This is a categorical field with 5 categories listed. We need to make sure only these 5 categories are allowed.

Tests New

This column has negative values in it and We need to make sure these negative values are taken care since tests cannot be negative.

Hospitalized New

This column has negative value, and we will be imposing constraint on this column.

Schools Closing

This is a Boolean field which says if the schools are closed or opened in the respective counties. And we will be imposing a constraint that this field remains Boolean.

There are columns like mentioned above which has negative values and all of them need to have constraints.

Merging Location and Covid data

We have two datasets where one has region codes and the other have county details based on the region code. Based on the region code we will be getting the details of the counties

Logging Process

The status of every row that would be moved out of the Staging area into the Master table would be logged in the logging table. With this, we can be assured that we are able to detect any errors that arise, and we can also perform data validation through row counts, if need be, to ensure all data is inserted.

Monthly Loads

We plan to load the entire data up to April 2020 into the data warehouse initially, and then add subsequent data with the help of Lookup function and a Conditional Split after the Staging area to load the additional data without any errors, primarily searching for duplicates, outliers and formatting.

Data Marts

We decided on 2 data marts (Pre-Covid and Post-Covid) that would help us answer some good questions about the data. From those data marts, we can perform further aggregations to answer more specific questions.

Data Validation

We would perform Data Validation primarily through the Logging Table, by checking the number of rows inserted. Also, for each row, we would be checking for formatting and data type errors that are bound to arise since we have multiple data sources.

Analytics Design

For the Star Schema, we have wanted to have information about the date and time, and one event ID for each of the 3 data sources. We would then branch out to the 3 different data sources.

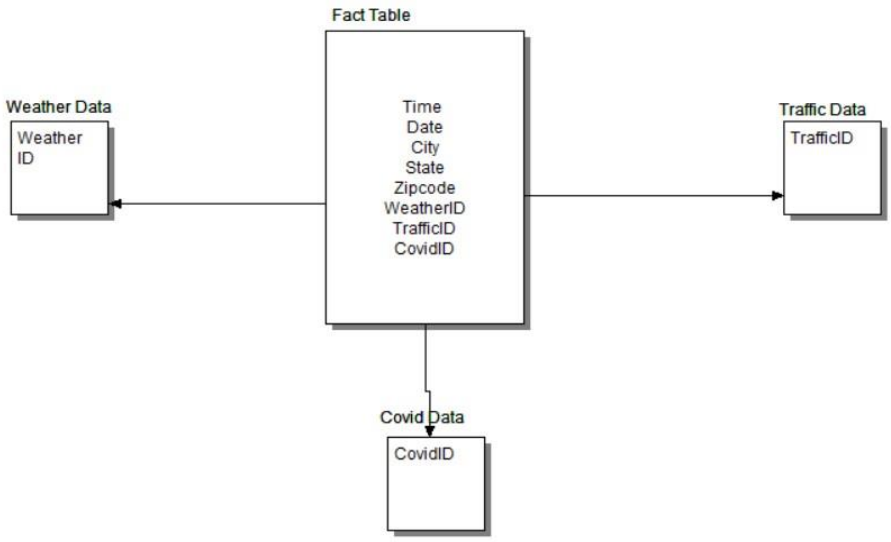


Fig. 2 Star Schema of tentative data warehouse

This is tentative and we plan to refine it even more. From this schema, we hope to build multiple OLAP cubes in SSAS or other supporting software that would help us understand the relationship between different attributes of the 3 data sources.

For reporting, we plan to use Tableau, as it is a powerful tool that will help us create meaningful and insightful visualizations.

Analytics Outcomes

We hope to understand the effects of weather on traffic in general, and how those effects were affected due to Covid. Also, what the impact of lockdown has been on the weather, and if there are any significant changes in the weather events because of covid.

Some questions about the data include:

What type of events have caused the most accidents?

At what time (season, weekday, etc.) for a particular weather condition, is the most likely for an accident to take place?

What effects did the Coronavirus have on traffic and accidents (whether they went down, increased due to stress, etc.)?

What impact did the Coronavirus have on the environment, and if so, in what form is it manifesting?

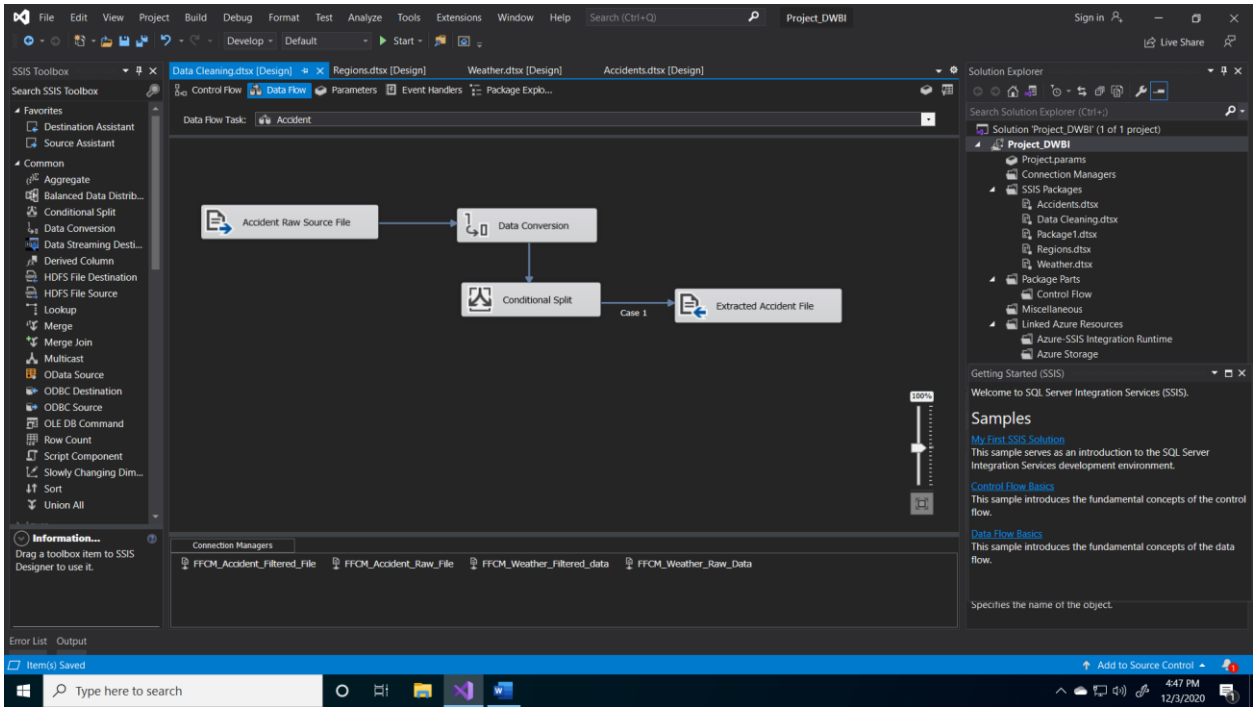
Design Document

As mentioned in the project proposal, we have 3 different data sources that we want to integrate into a single fact table, from which we can perform relevant analysis.

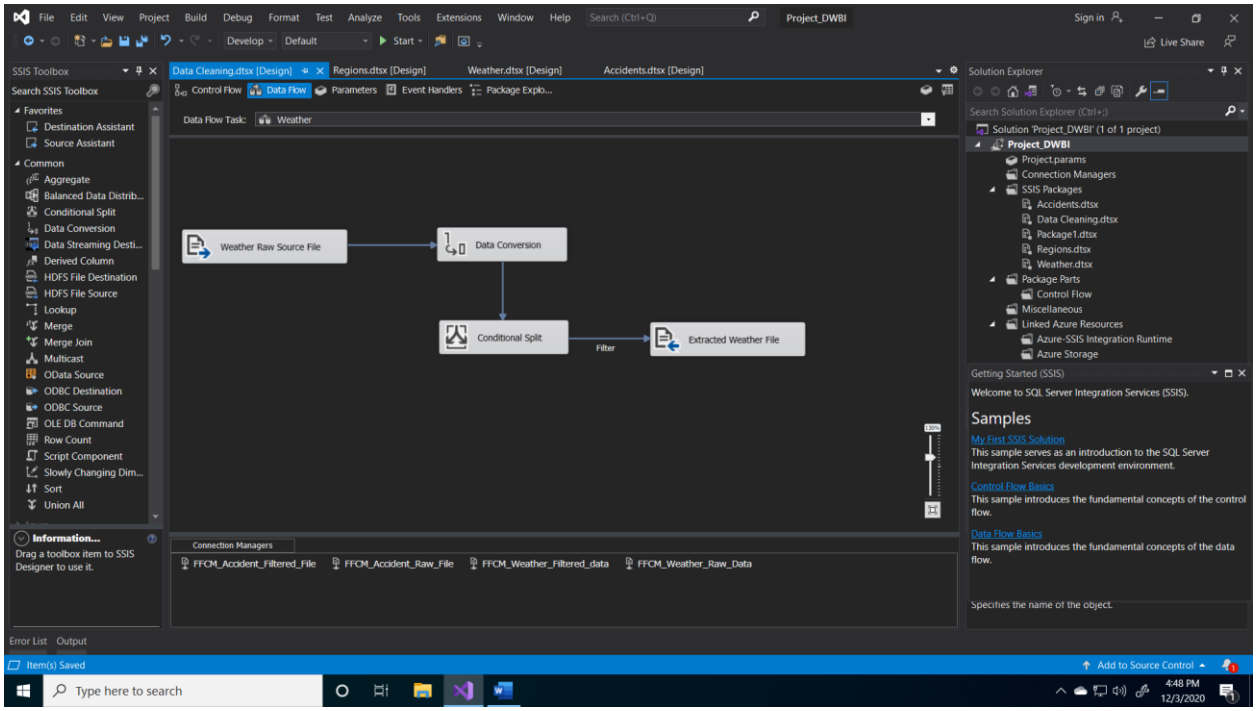
Since the raw data that was obtained is not in the proper form, we used SSIS to clean and extract the relevant information by removal of unnecessary columns and information.

Initial Data Cleaning

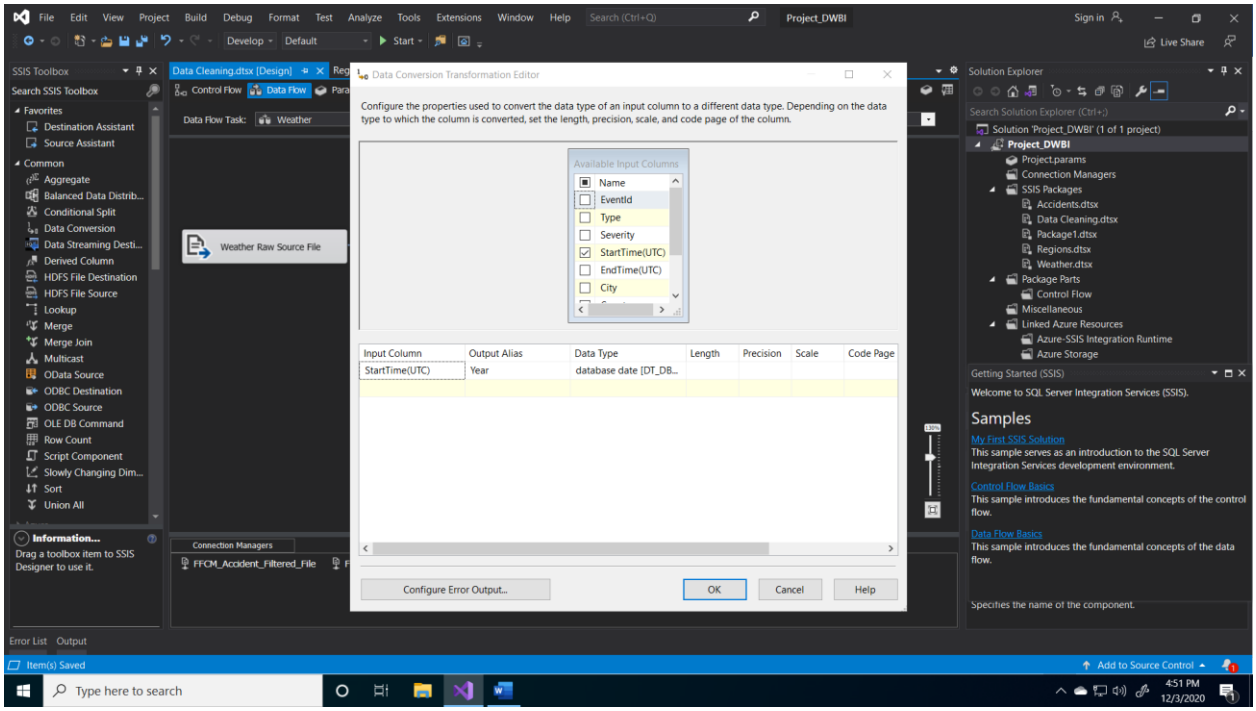
Data extraction Flow for Accident Data from raw source file to cleaned file



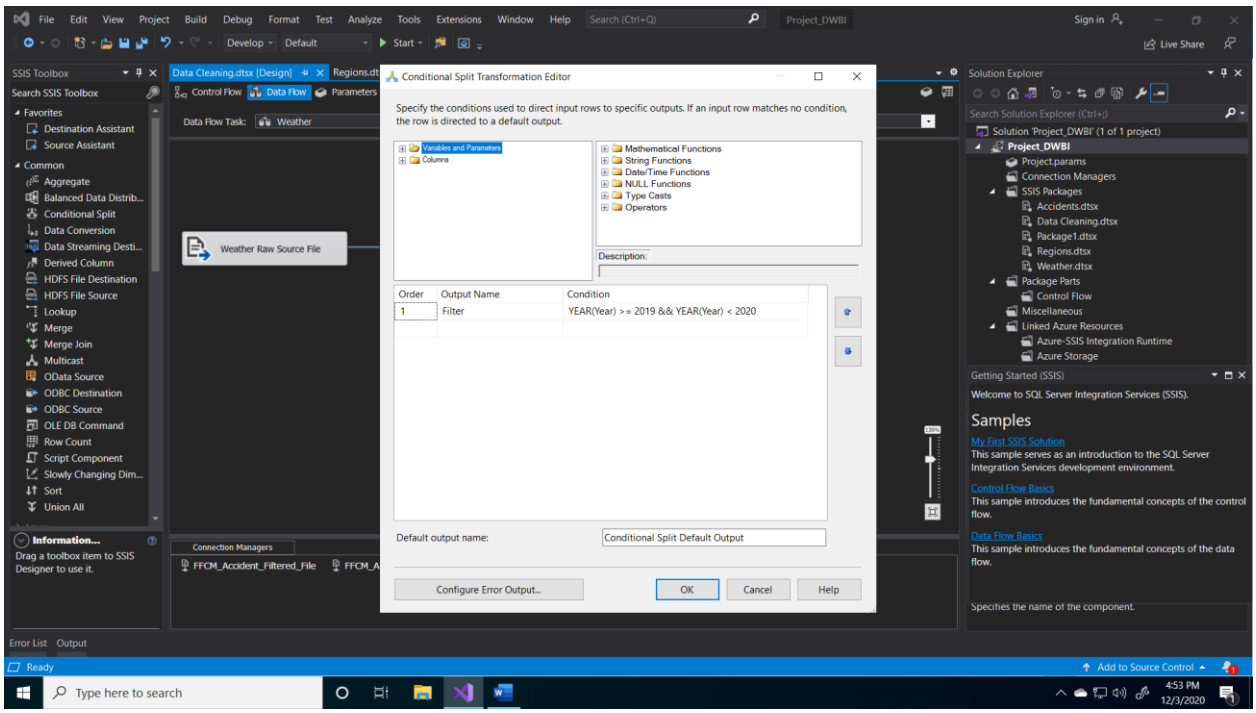
Data extraction Flow for Weather source file to cleaned file



For both Accident and Weather files we extracted Start Date value from UTC format of given StartDate, as information about time is not relevant for Covid data.

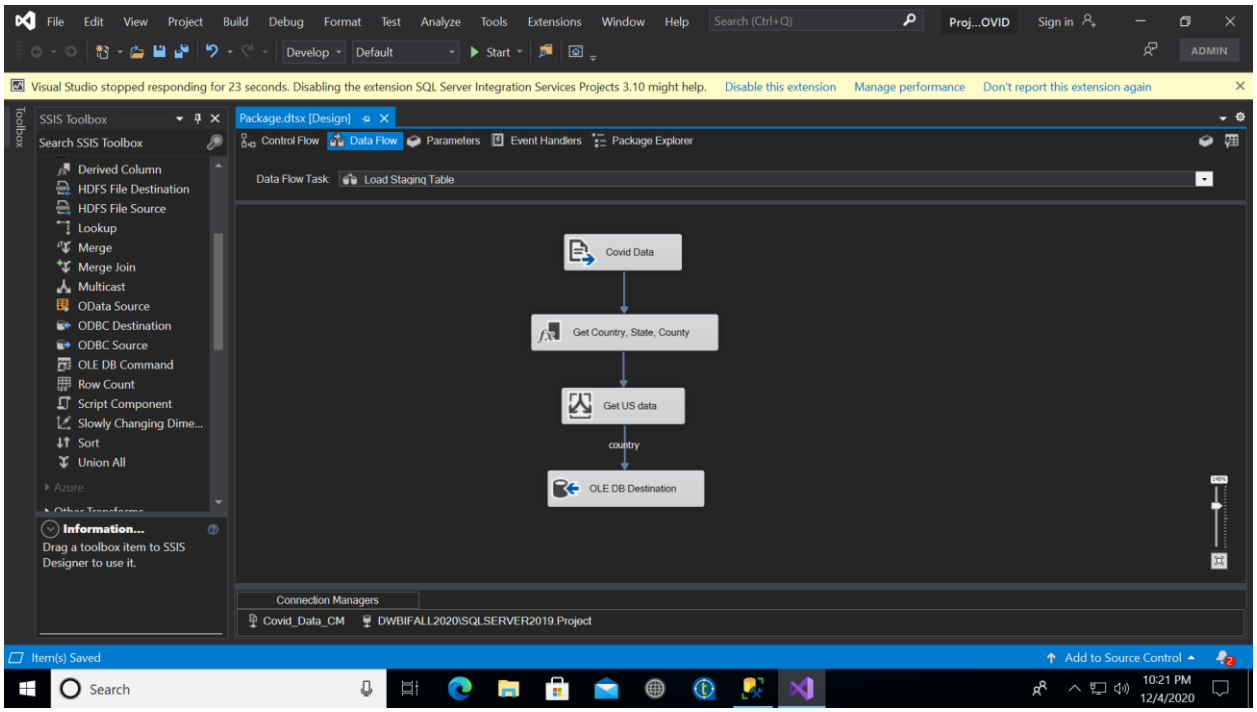


Then, using conditional split, we extracted year value using Date/Time Function and added desired value.



The output from the conditional split is then stored into a csv file.

Data extraction Flow for Covid Data from raw source file to cleaned file. The covid data file has all the countries data over the world and we only want US data. Hence extracting only US country data for our project.



Loading Data into Staging Tables

Weather Data:

Our weather data consists of the following columns,

EventId: A unique value represents a weather event

Type: Specifies what type of weather.

Severity: Indicates the severity of given weather event.

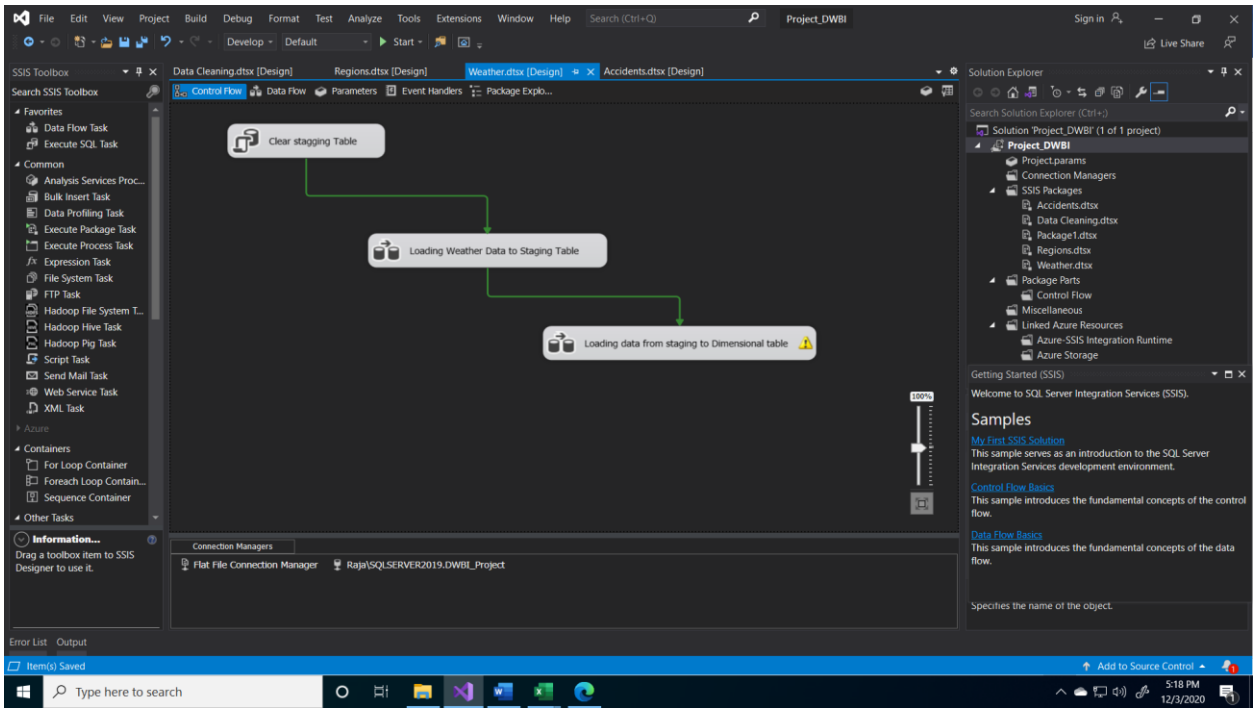
StartTime (UTC): The StartTime of the weather event in UTC format.

EndTime (UTC): The EndTime of the weather event in UTC format.

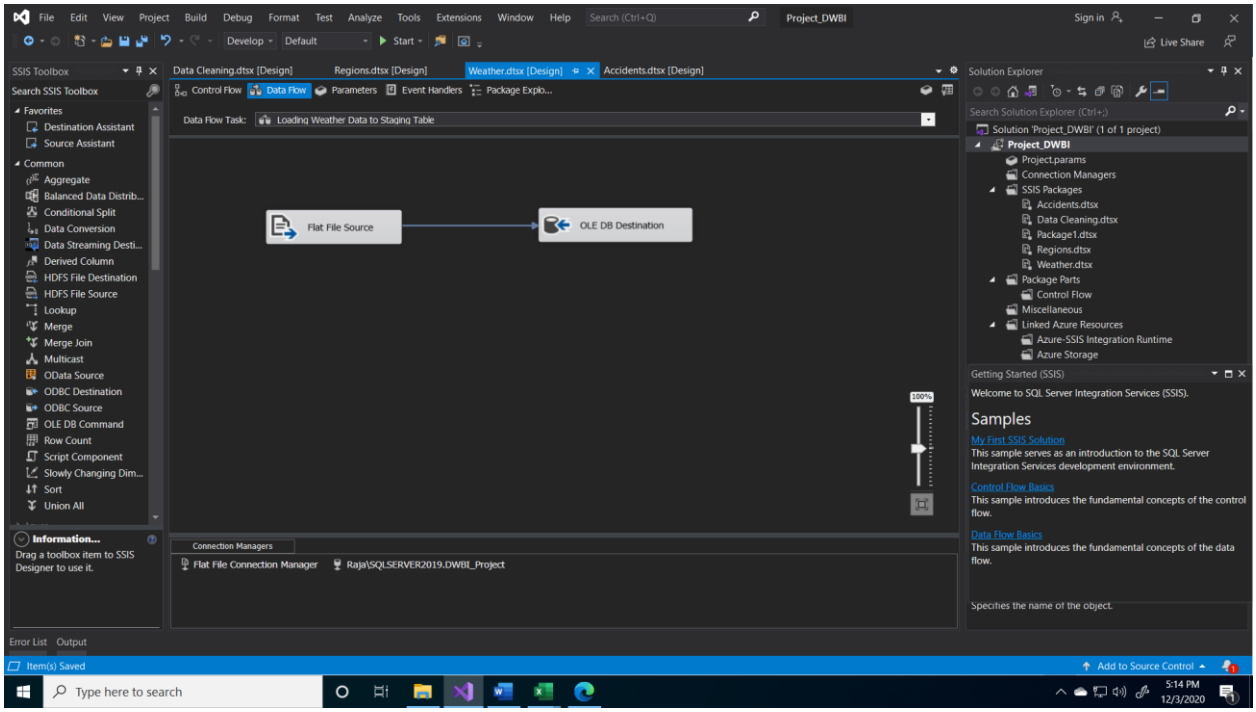
City, County, State and ZipCode: Provides geographical information of weather event.

TimeZone, LocationLat, LocationLng, AirportCode: We have removed these columns.

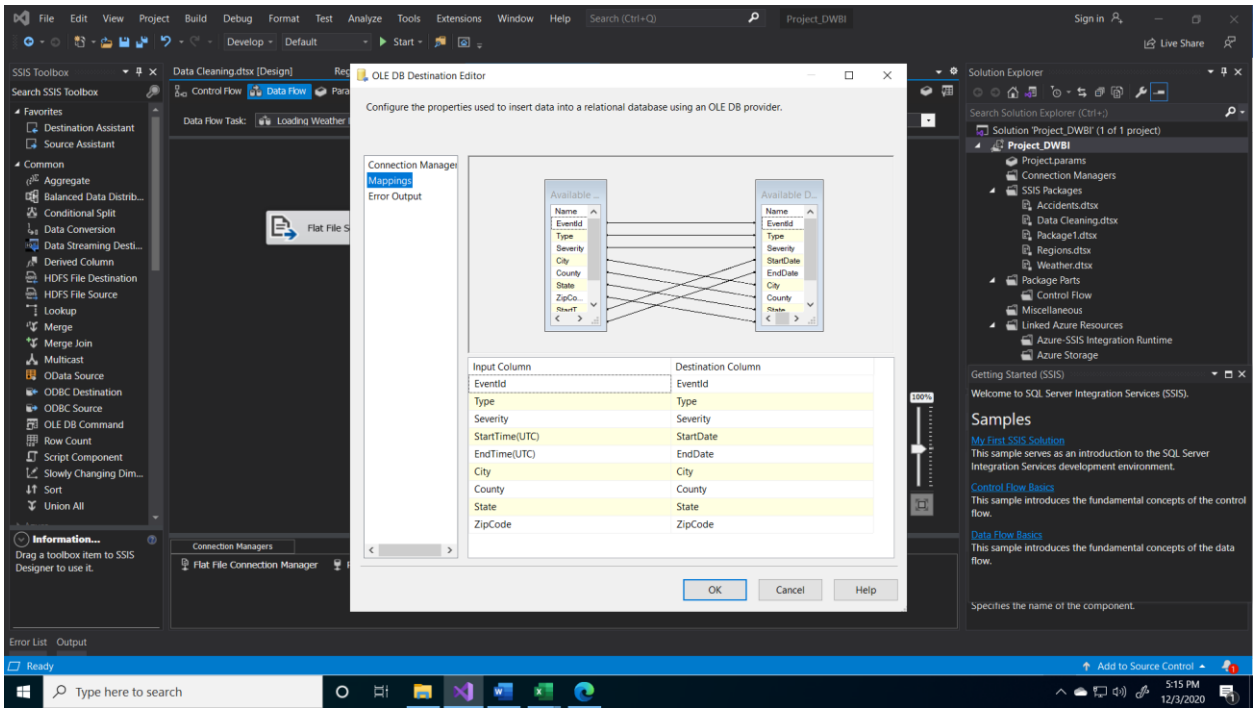
First, we clear staging table to remove all existing values in the table



Loading data from Source to staging table.



All required columns are directly loaded to staging table having columns Vchar Data Type



Accident/Traffic Data:

Our Accident data consists of the following columns,

EventId: A unique value represents an accident

Type: Specifies what type of Accident.

Description: Provides a brief Description of the Accident/Traffic event.

Severity: Indicates the severity of given Accident.

StartTime (UTC): The StartTime of the Accident/Traffic in UTC format.

EndTime (UTC): The EndTime of the Accident/Traffic in UTC format.

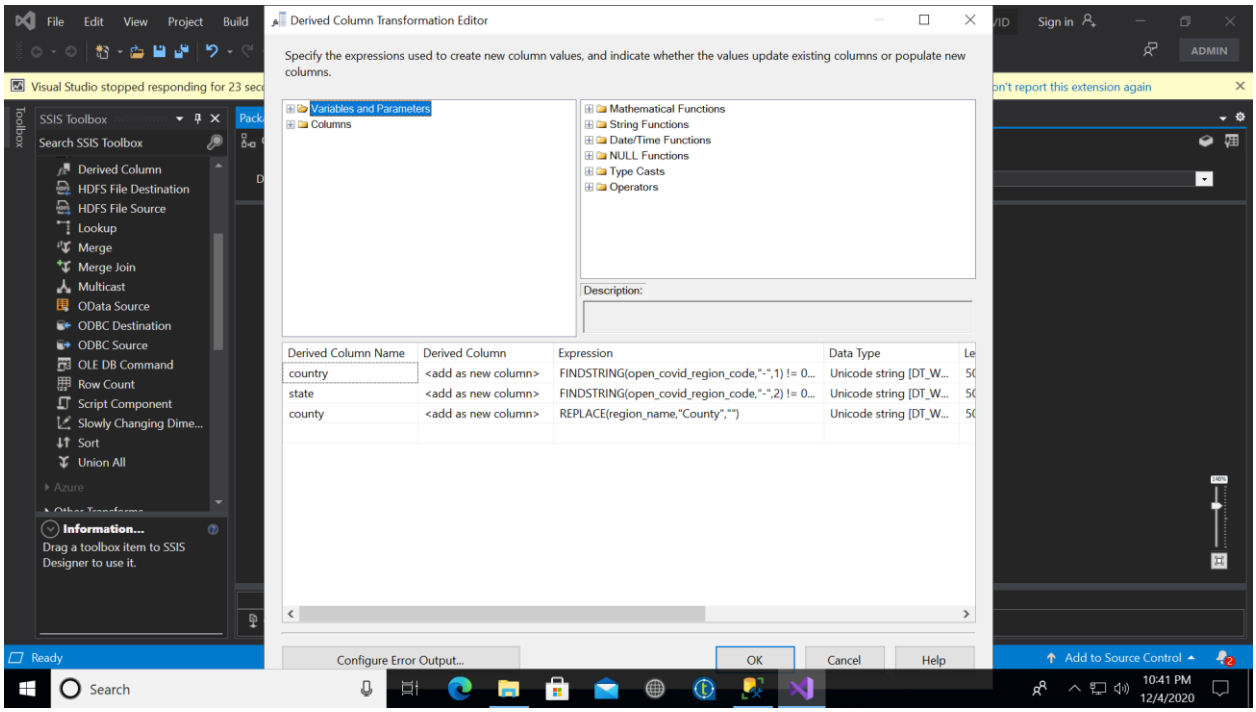
City, County, State and ZipCode: Provides geographical information of Accident/Traffic.

TimeZone, LocationLat, LocationLng, AirportCode, TMC, Distance (Mi), Number, Street, Side: We have removed these columns.

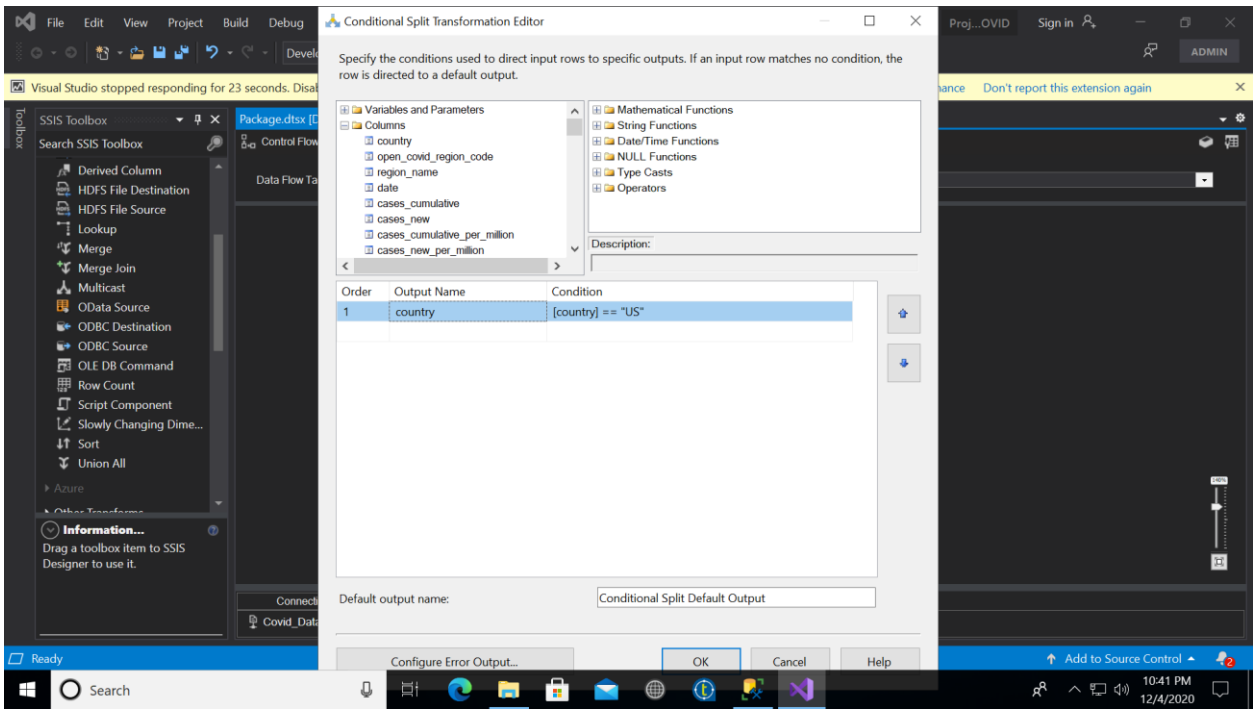
The steps followed are like that of Weather Data.

Covid data:

We have a column ‘open_covid_region_code’ which has got values like ‘US-WA-53061’, extracting country and state details from this field using derived columns.



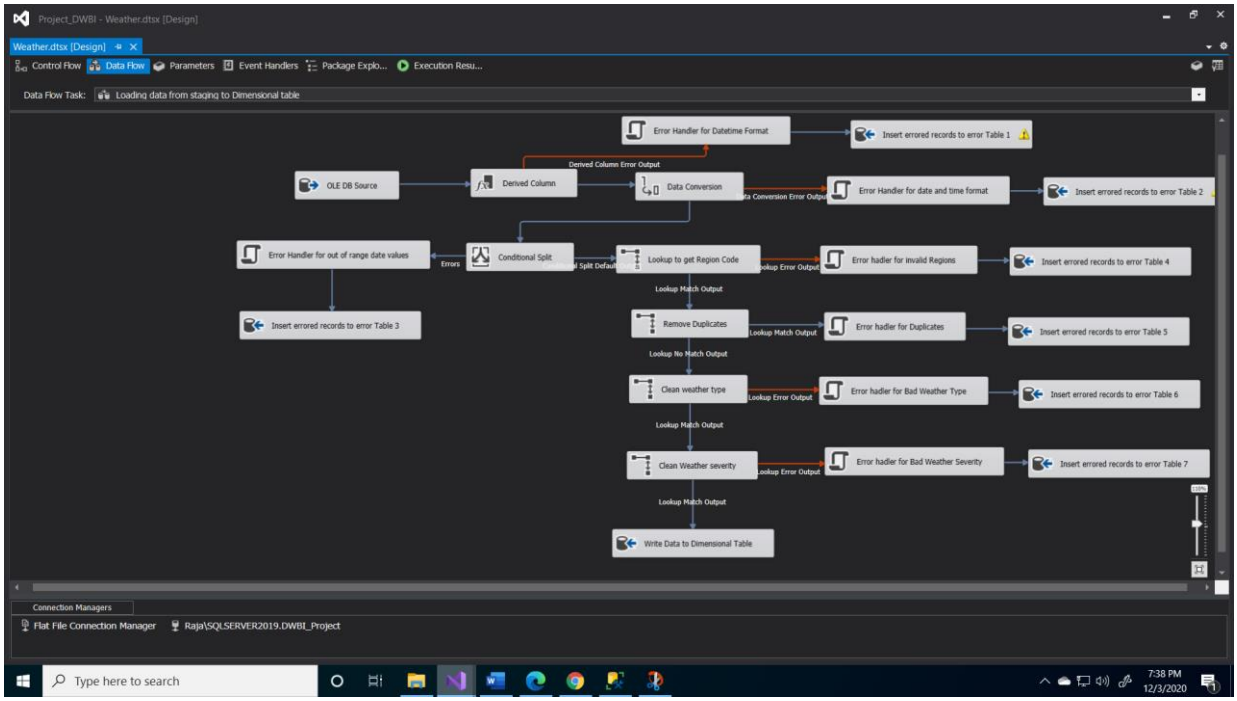
After extracting country from the field, we are loading only US data into the staging table using conditional split.



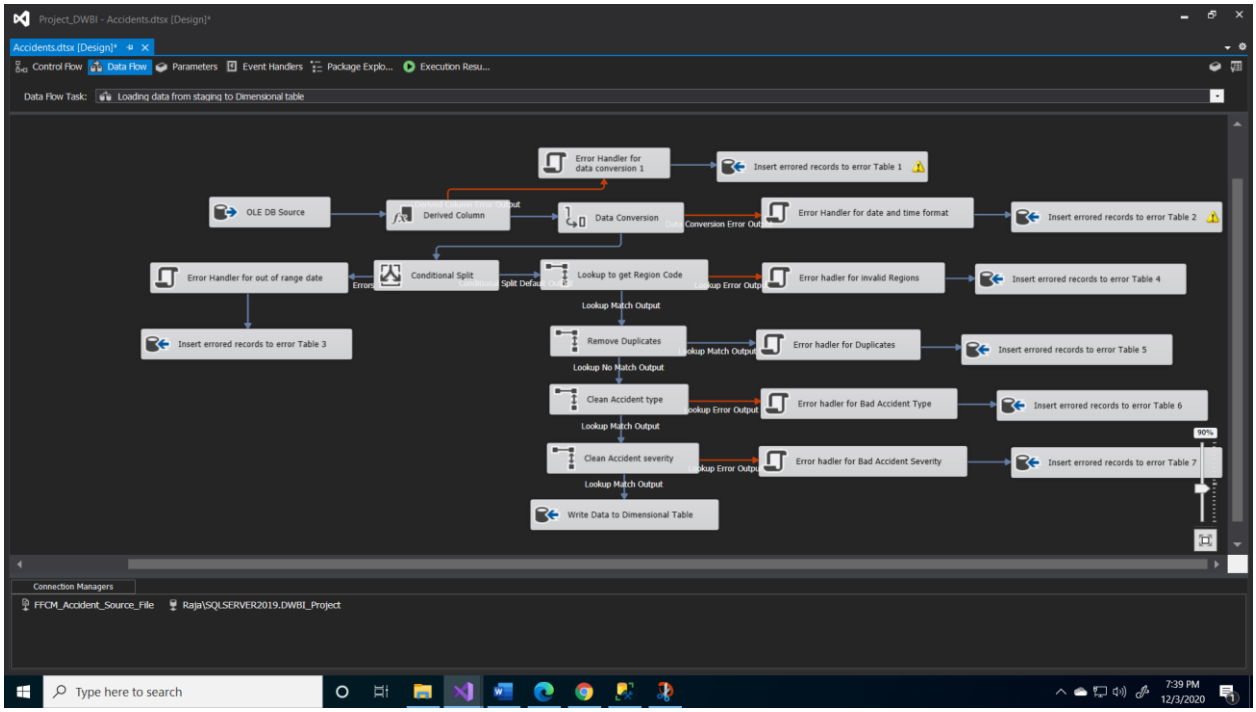
Loading data from Staging to Dimensional

Next, we insert the data from staging to dimensional table after performing required transformation.

Weather Data



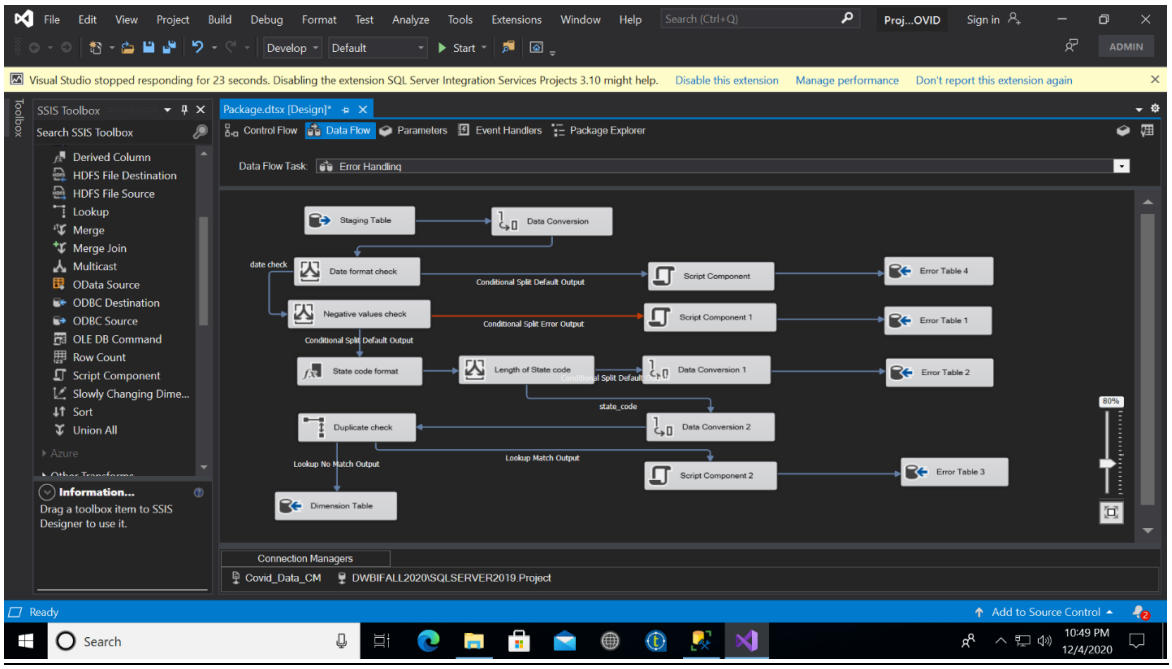
Accident Data



We can observe that there are millions of records that contain the same information about County, City, State and Zipcode.

Hence, to avoid this redundancy in the dimensional table, we created an Outrigger for regions that contains all the Geographic Information, enabling us to store only the region code in the dimension table.

Covid data



Dimensional Table for Regions:

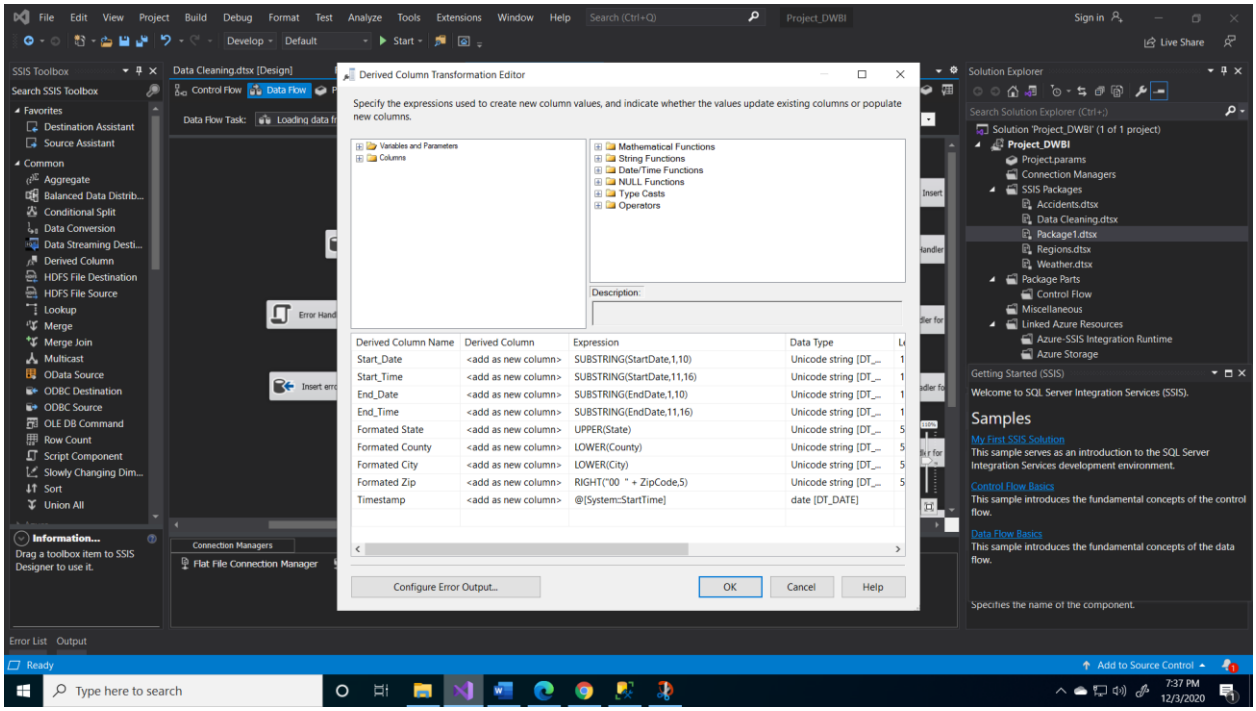
Snippet of data loaded in Regions table, here Region_Code is the Surrogate key, which acts as Foreign Key for other table by uniquely matching with Zipcode, State code, City and County detail.

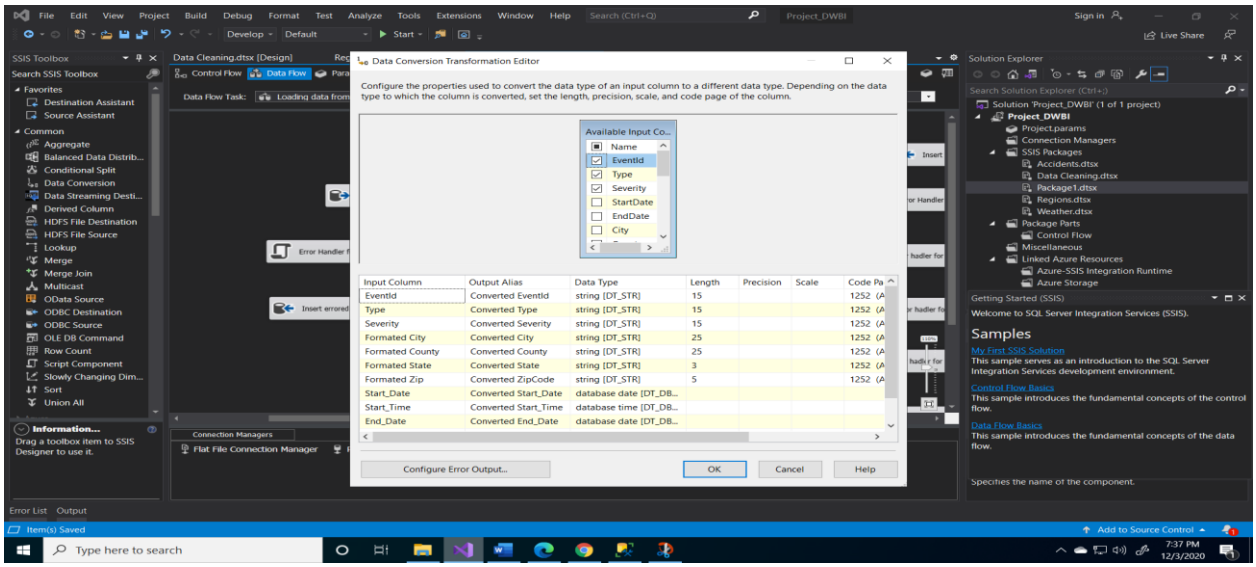
	Region_Code	zipcode	State	State_Code	County_Name	City_Name	TimeStamp
1	10131	01001	massachusetts	MA	hampden	agawam	2020-12-01 16:04:17.000
2	10132	01002	massachusetts	MA	hampshire	amherst	2020-12-01 16:04:17.000
3	10133	01003	massachusetts	MA	hampshire	amherst	2020-12-01 16:04:17.000
4	10134	01005	massachusetts	MA	worcester	barre	2020-12-01 16:04:17.000
5	10135	01007	massachusetts	MA	hampshire	belchertown	2020-12-01 16:04:17.000
6	10136	01008	massachusetts	MA	hampden	blandford	2020-12-01 16:04:17.000
7	10137	01009	massachusetts	MA	hampden	bondsville	2020-12-01 16:04:17.000
8	10138	01010	massachusetts	MA	hampden	brimfield	2020-12-01 16:04:17.000
9	10139	01011	massachusetts	MA	hampden	chester	2020-12-01 16:04:17.000
10	10140	01012	massachusetts	MA	hampshire	chesterfield	2020-12-01 16:04:17.000
11	10141	01013	massachusetts	MA	hampden	chicopee	2020-12-01 16:04:17.000
12	10142	01020	massachusetts	MA	hampden	chicopee	2020-12-01 16:04:17.000
13	10143	01022	massachusetts	MA	hampden	chicopee	2020-12-01 16:04:17.000
14	10144	01026	massachusetts	MA	hampshire	cummington	2020-12-01 16:04:17.000
15	10145	01027	massachusetts	MA	hampshire	easthampton	2020-12-01 16:04:17.000

Error Handling for Weather and Accident/Traffic Data:

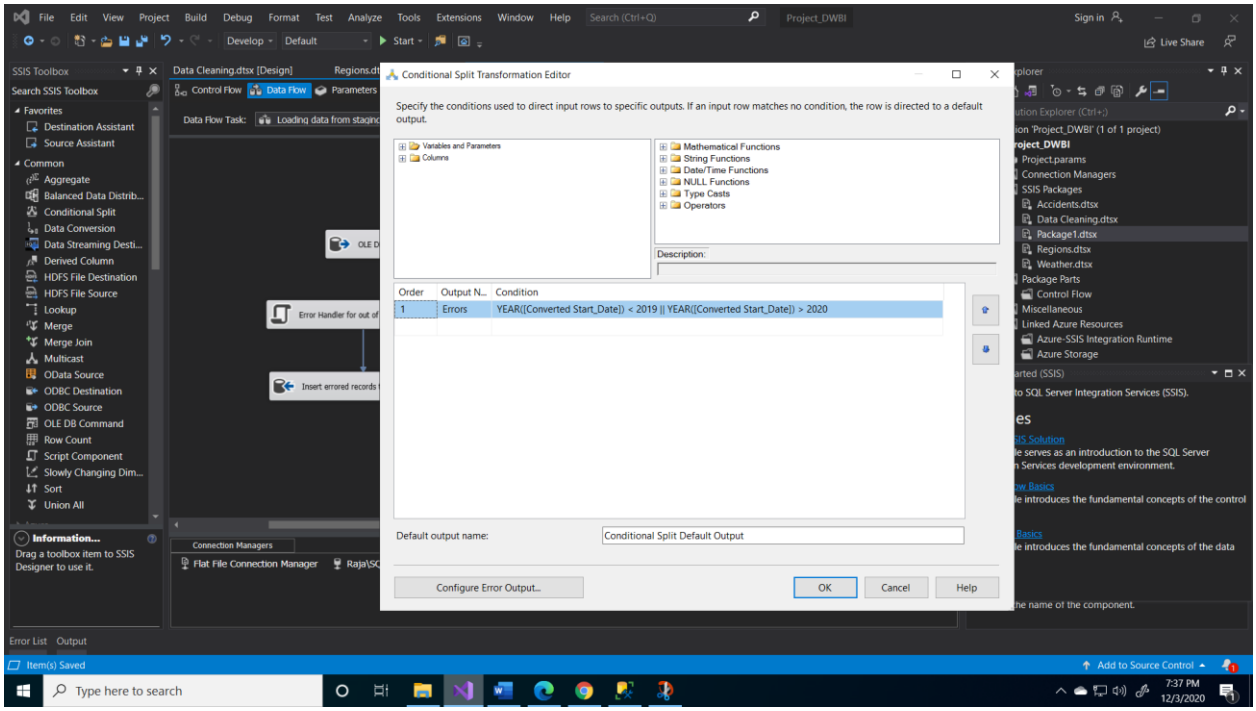
All Types of error records are stored in Error table, with Failure Reason and a brief Failure Description.

1. Check if the incoming record as proper start/ End Time formats.
   By making use of derived column to separate Start and End Date into Time and Date. We perform the required check if it gets converted into valid value correctly or not. Records which fail the conversion are sent to error table.



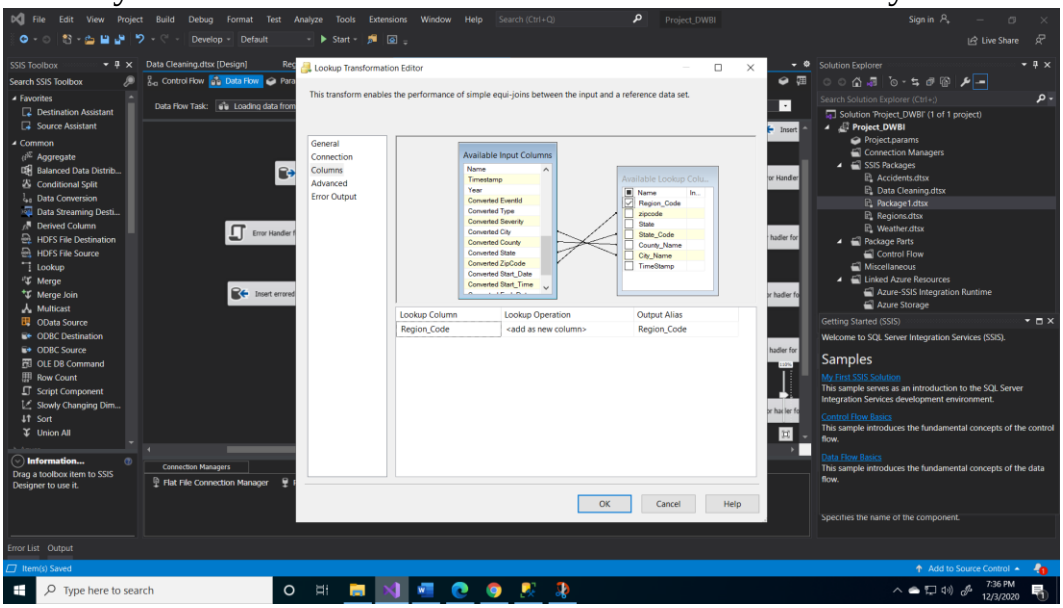


2. If the Date is within the desired range (year between 2019 and 2020), by utilizing the Conditional split we check if an incoming record is within specified range. Only the record which satisfies this condition are sent for further processing. While remaining data is written into error table.

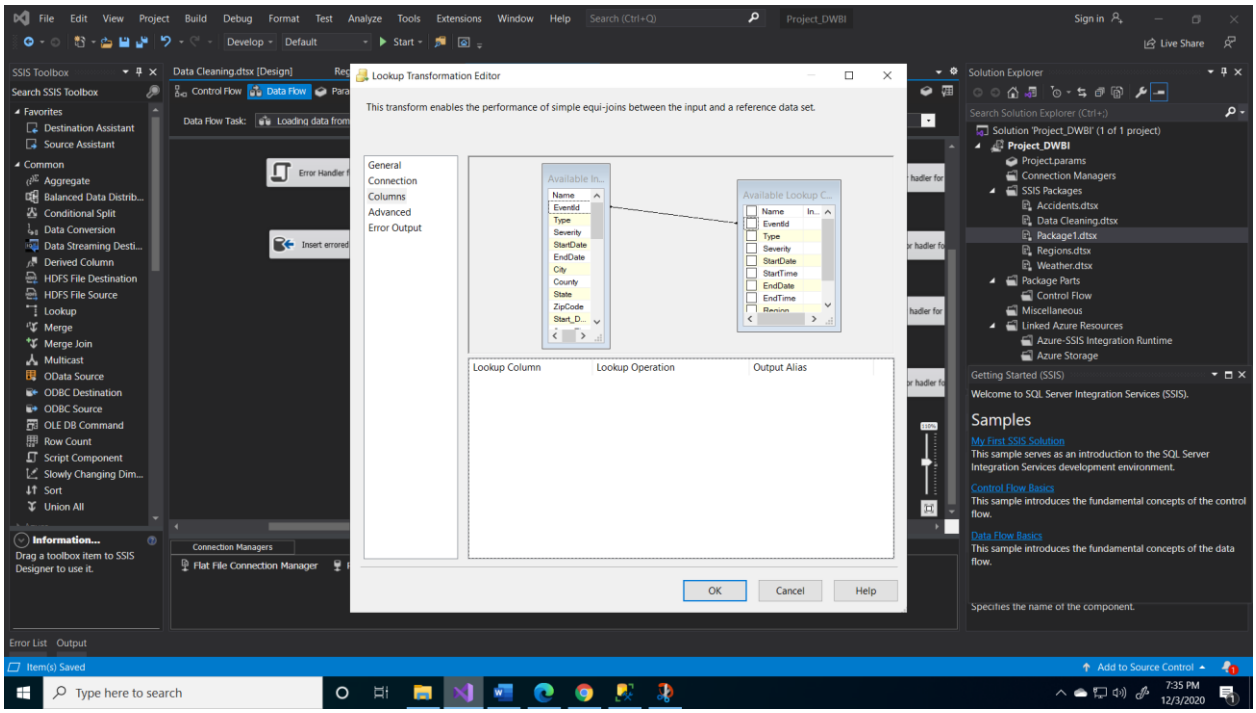


3. Eliminate records with improper Zipcode, City, County and State Code values. As mentioned, before we have Implemented a Table which holds a unique value for each combination of geographical information. This is performed with a help of a lookup table.
   We are converting all incoming geographical location to lower case for city, County, and upper case for State Code, Zipcode is made sure it's in the proper 5-digit format.

   For All matching records, the surrogate key from Regions table is sent in as a way to normalize the tables and reduce data redundancy.

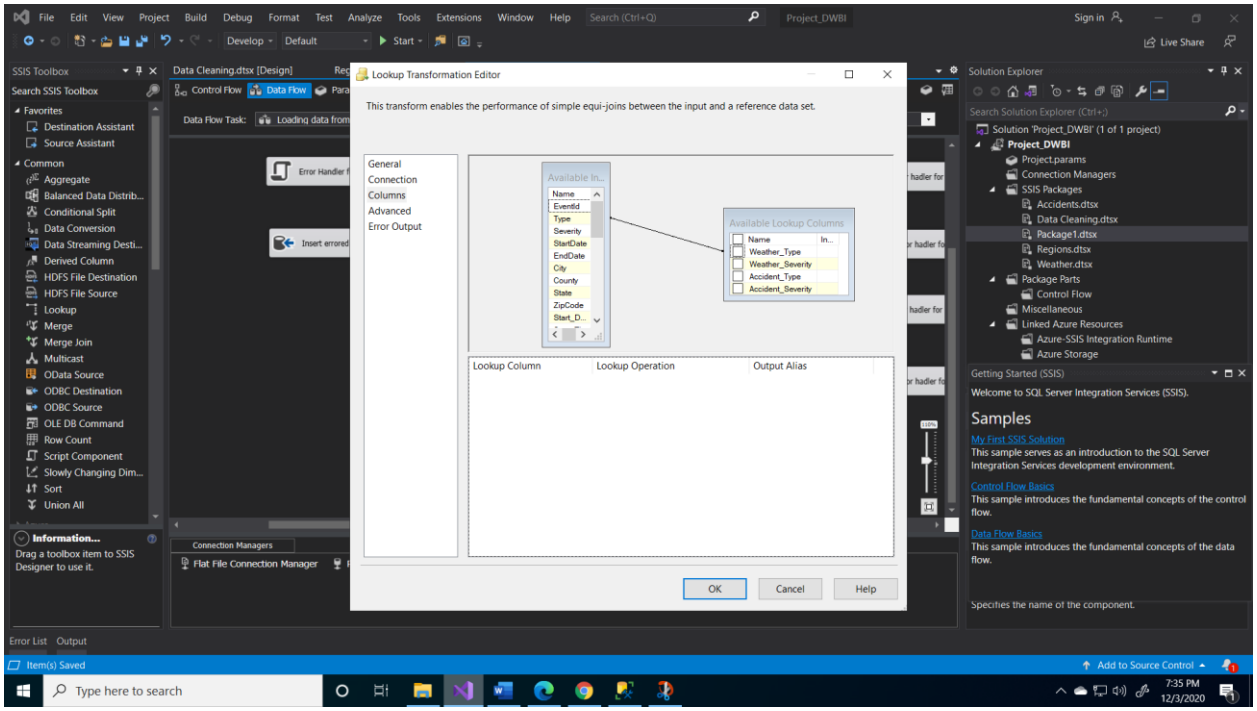


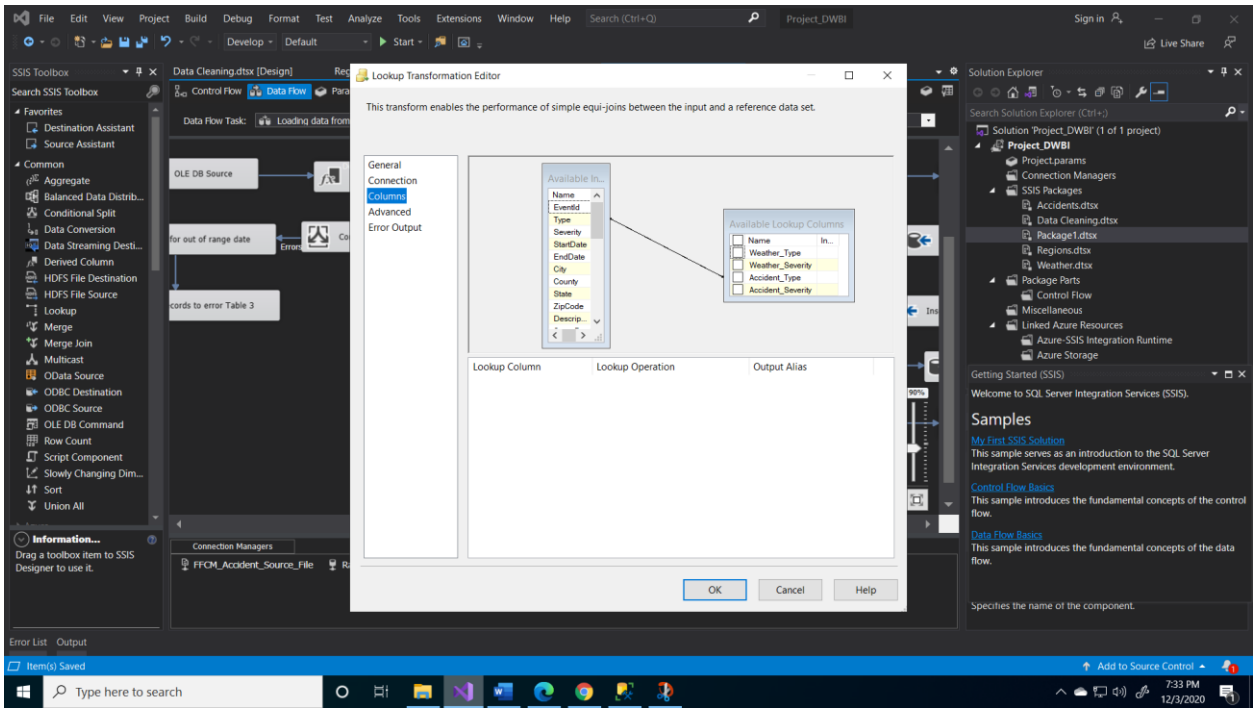
4. Check for duplicate entries.
   As each Traffic/Accident or Weather event is uniquely identified by EventID we have added a Lookup table to eliminate loading same data again into Dimensional table.
   It Checks if the incoming record with given EventID is already present, if it's already an existing record it is sent to the error table. While new entries are processed further.



5. Check For invalid Type of weather and Accident/Traffic Event.
   We have a set of predefined Type for Weather and Accident/Traffic. We have stored these values in a table called Standard values, if the incoming record as any value which is not present in the below list, they're considered error or out of scope values. And they're sent to the error table. This performed using a Lookup Table.

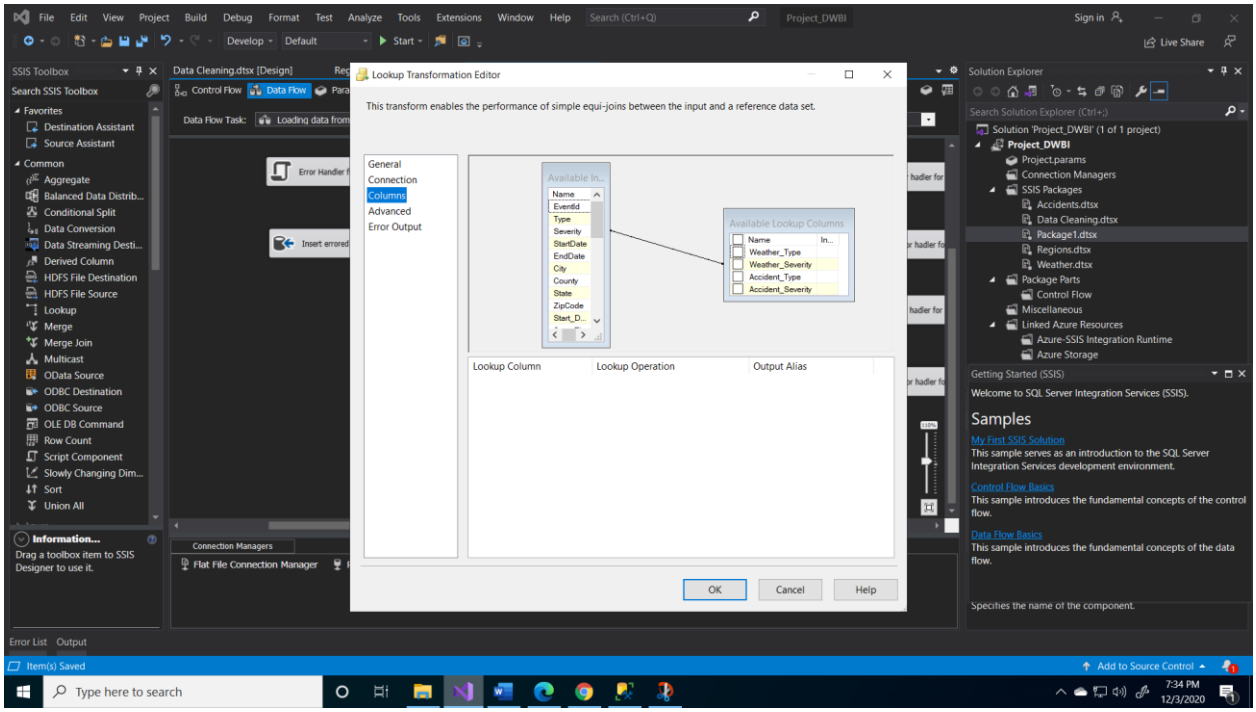
	Weather_Type	Accident_Type
1	Cold	Broken-Vehicle
2	Precipitation	Construction
3	Fog	Lane-Blocked
4	Snow	Event
5	Rain	Congestion
6	Hail	Flow-Incident
7	Storm	Accident

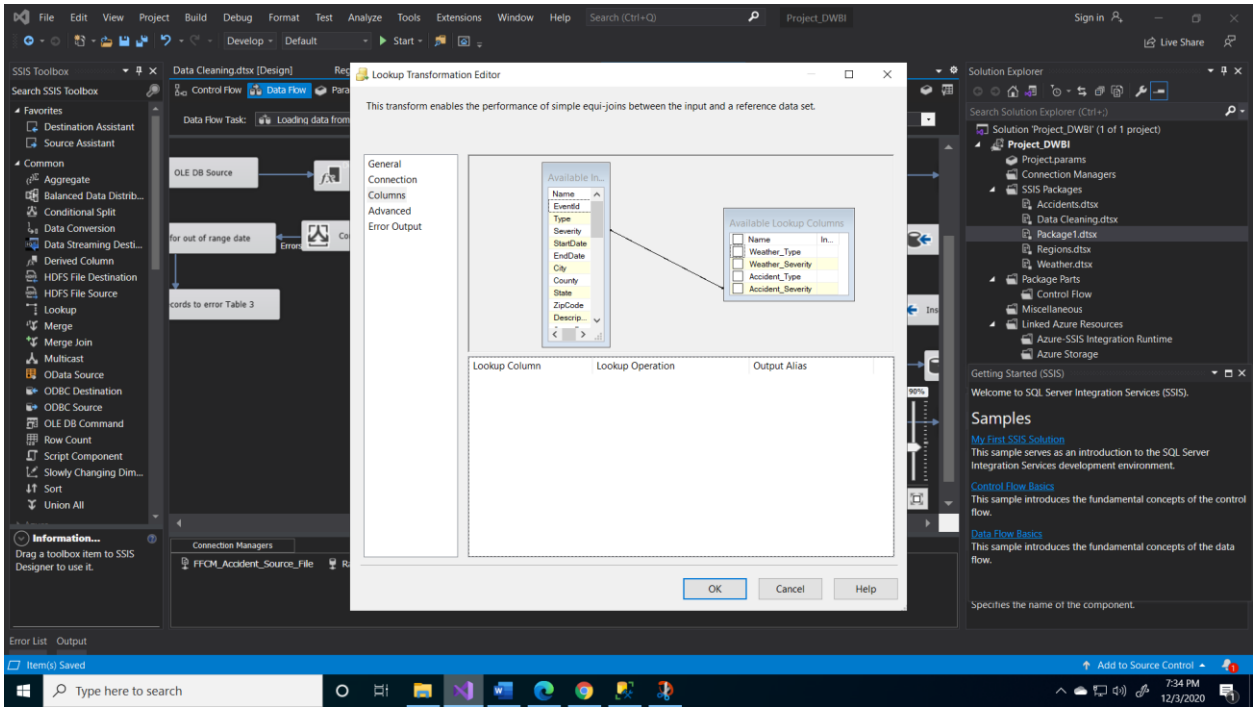

Lookup Transformation Editor
This transform enables the performance of simple equi-joins between the input and a reference data set.
General
Connection
Columns
Advanced
Error Output
Available In...
Available Lookup Columns
Lookup Column
Lookup Operation
Output Alias
OK
Cancel
Help
Data Cleaning.dtsx [Design]
Control Flow
Data Flow
Data Flow Task:
Loading data from
Connection Managers
Flat File Connection Manager
SSIS Toolbox
Solution Explorer
Project_DWBI
Getting Started (SSIS)
Samples
Information...
Drag a toolbox item to SSIS Designer to use it.
Specifies the name of the component.
Item(s) Saved
Add to Source Control
Type here to search
7:35 PM
12/3/2020


Lookup Transformation Editor
This transform enables the performance of simple equi-joins between the input and a reference data set.
General
Connection
Columns
Advanced
Error Output
Available In...
Available Lookup Columns
Lookup Column
Lookup Operation
Output Alias
OK
Cancel
Help
Data Cleaning.dtsx [Design]
Control Flow
Data Flow
Data Flow Task:
Loading data from
OLE DB Source
Connection Managers
FFCM_Accident_Source_File
SSIS Toolbox
Solution Explorer
Project_DWBI
Getting Started (SSIS)
Samples
Information...
Drag a toolbox item to SSIS Designer to use it.
Specifies the name of the component.
Item(s) Saved
Add to Source Control
Type here to search
7:33 PM
12/3/2020

6. Check for Severity of weather and Accident/Traffic Event
   Similarly, for the Severity of Weather or Accident/Traffic we have predefined values, which are stored standard values table. If Any record has values which are not from the list they're considered as out of scope or error. Again, we have implemented a Lookup table for this activity.

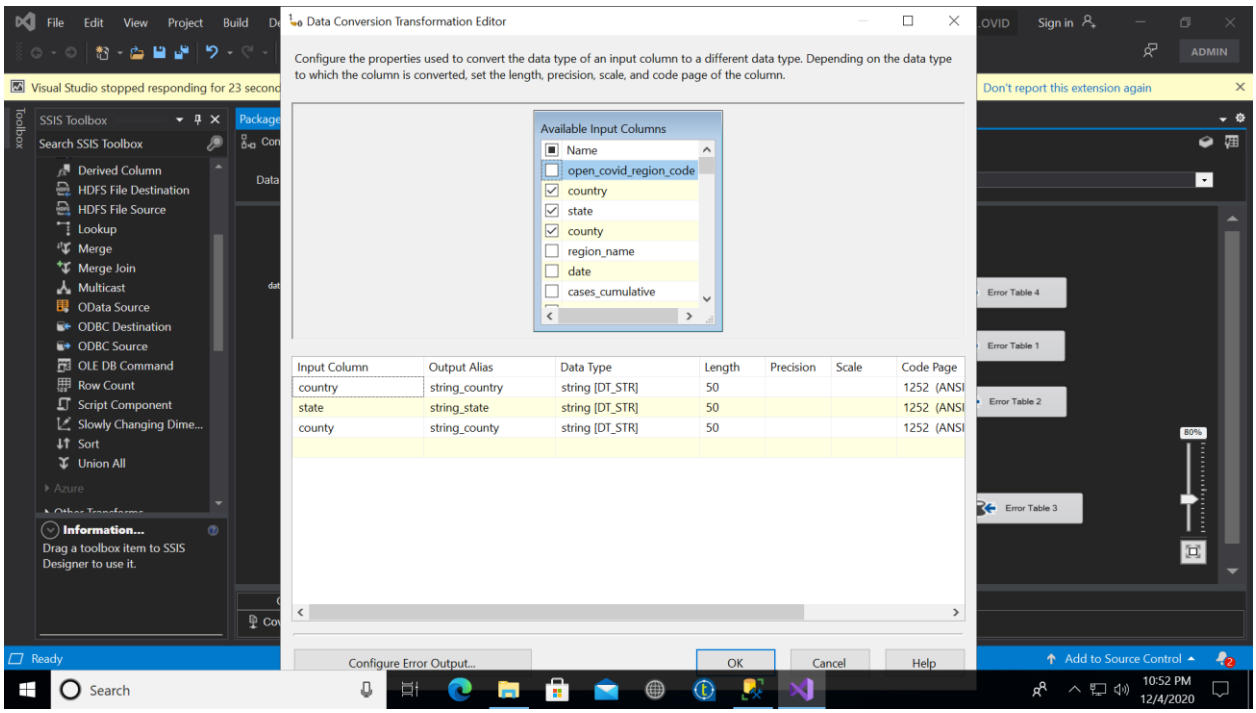
	Weather_Severity	Accident_Severity
1	Severe	Moderate
2	UNK	NULL
3	Heavy	Fast
4	Moderate	Other
5	Light	Long
6	Other	Short
7	NULL	Slow



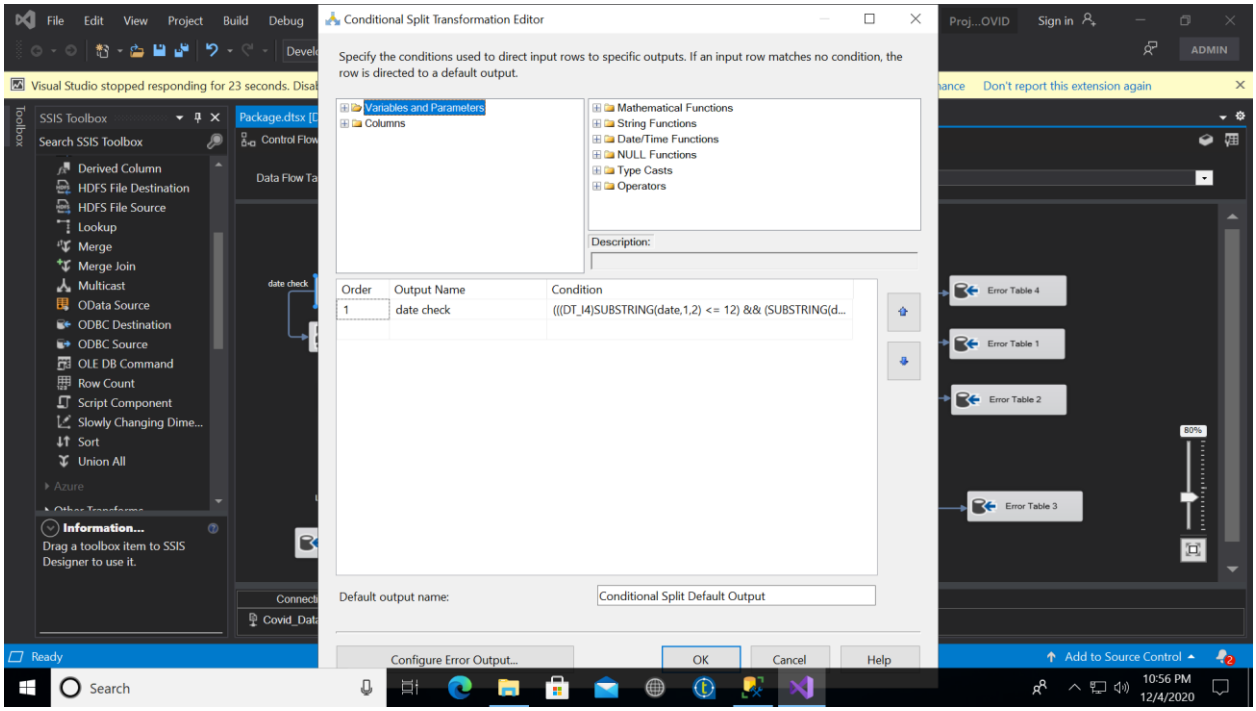


Error Handling for Covid Data

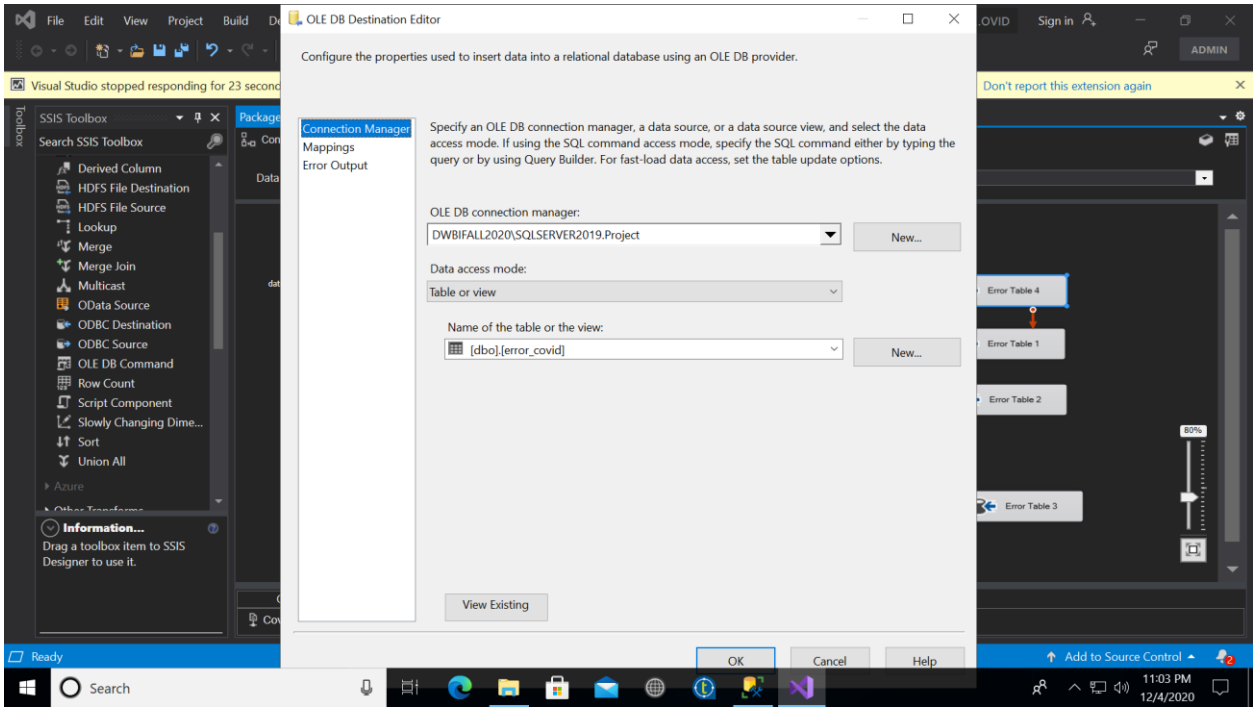
Converting fields Country, State, County into string using data conversion task.



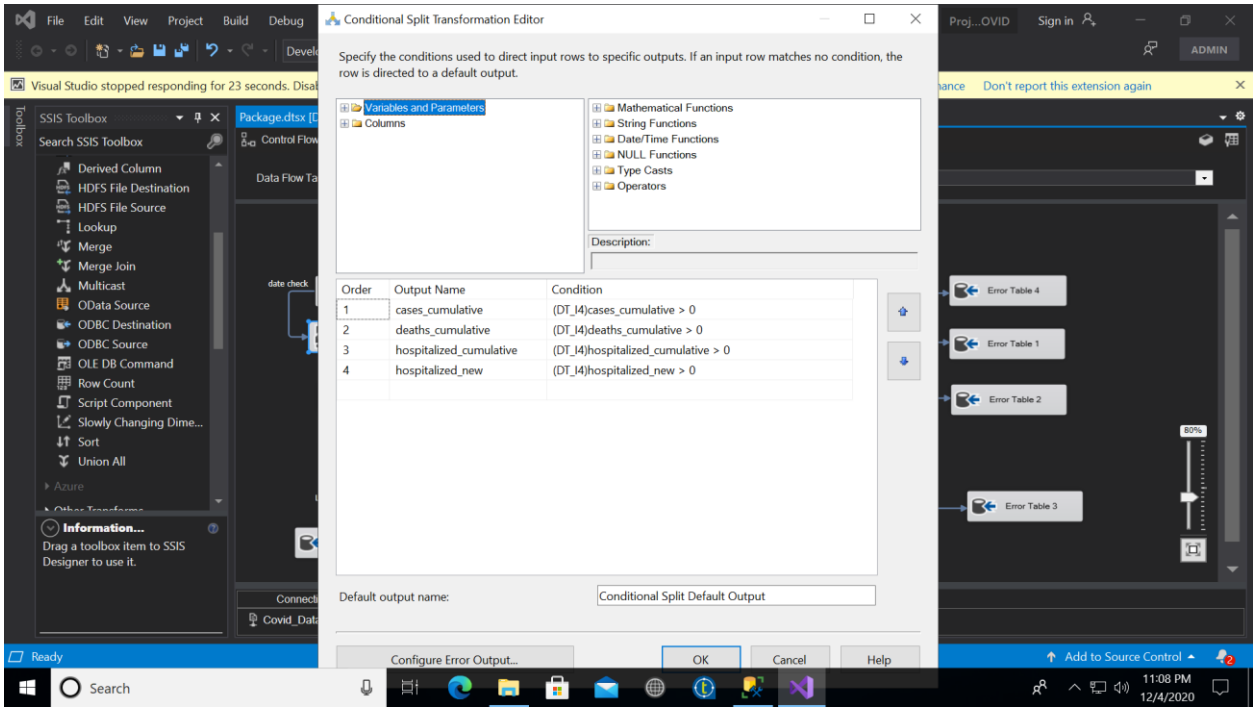
Performing Date format check: checking if the date format is correct .i.e.‘MM/DD/YYY’



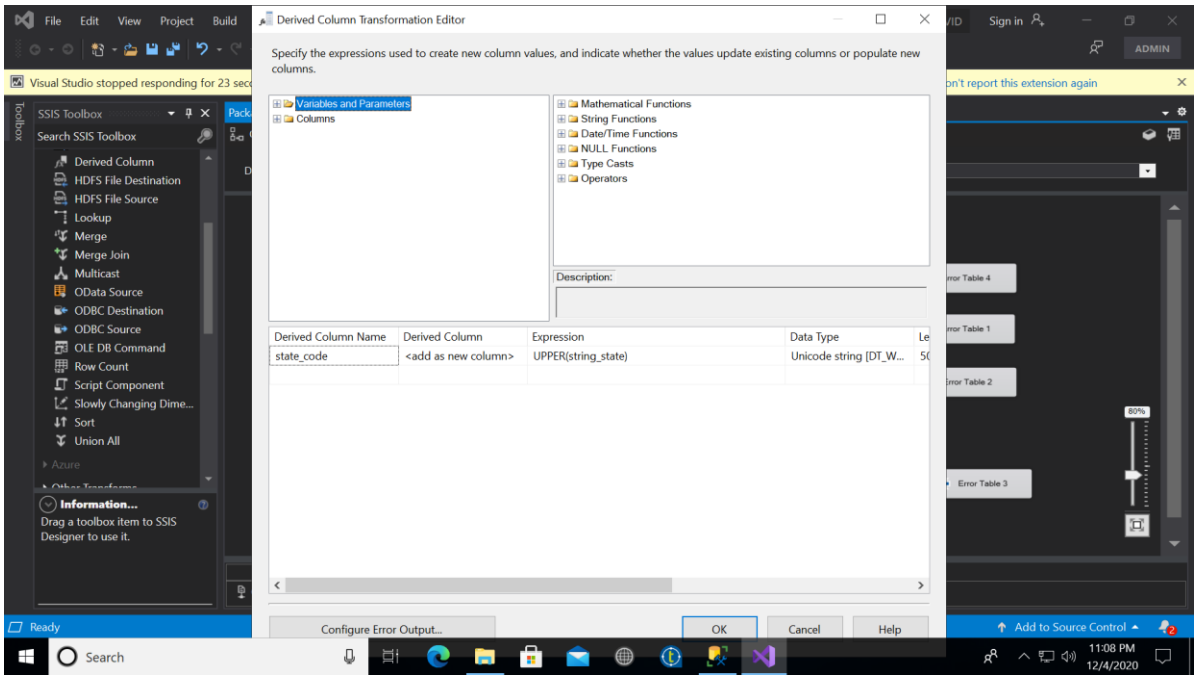
If the date is not in the standard format, then moving that record to Error table

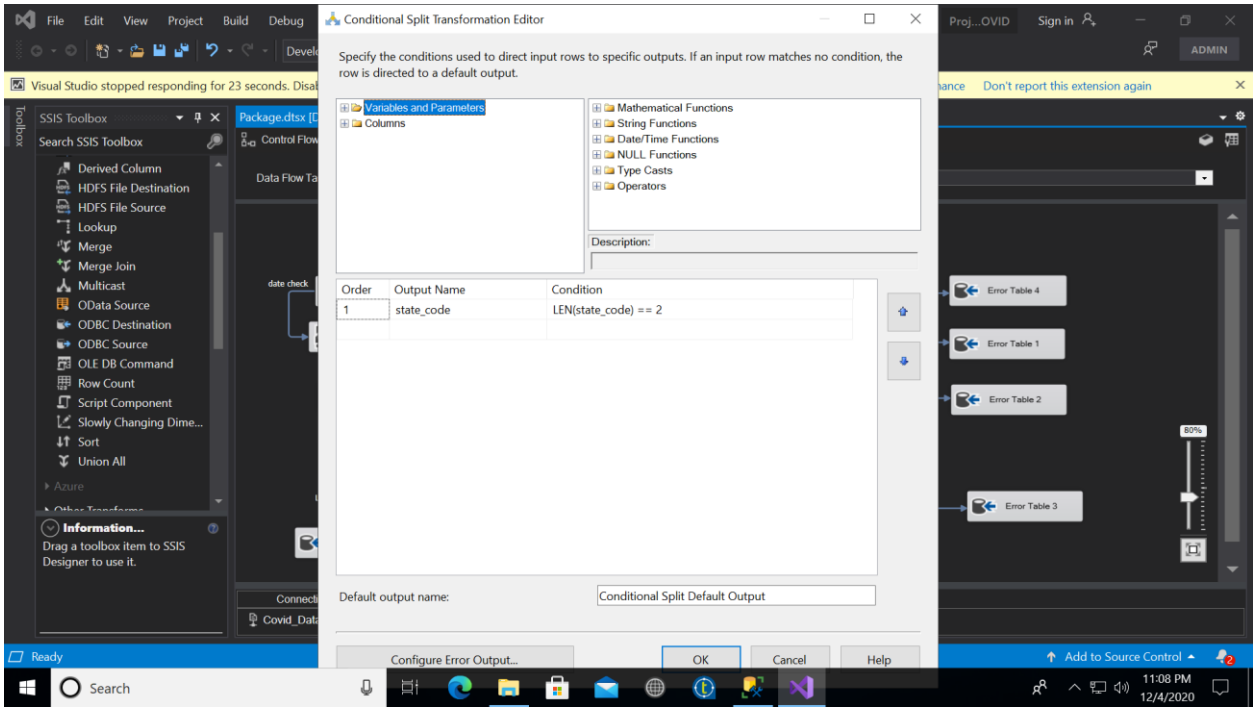


Negative values check: if any of the fields has negative value then moving the entry to the error table if not then moving to dimension table

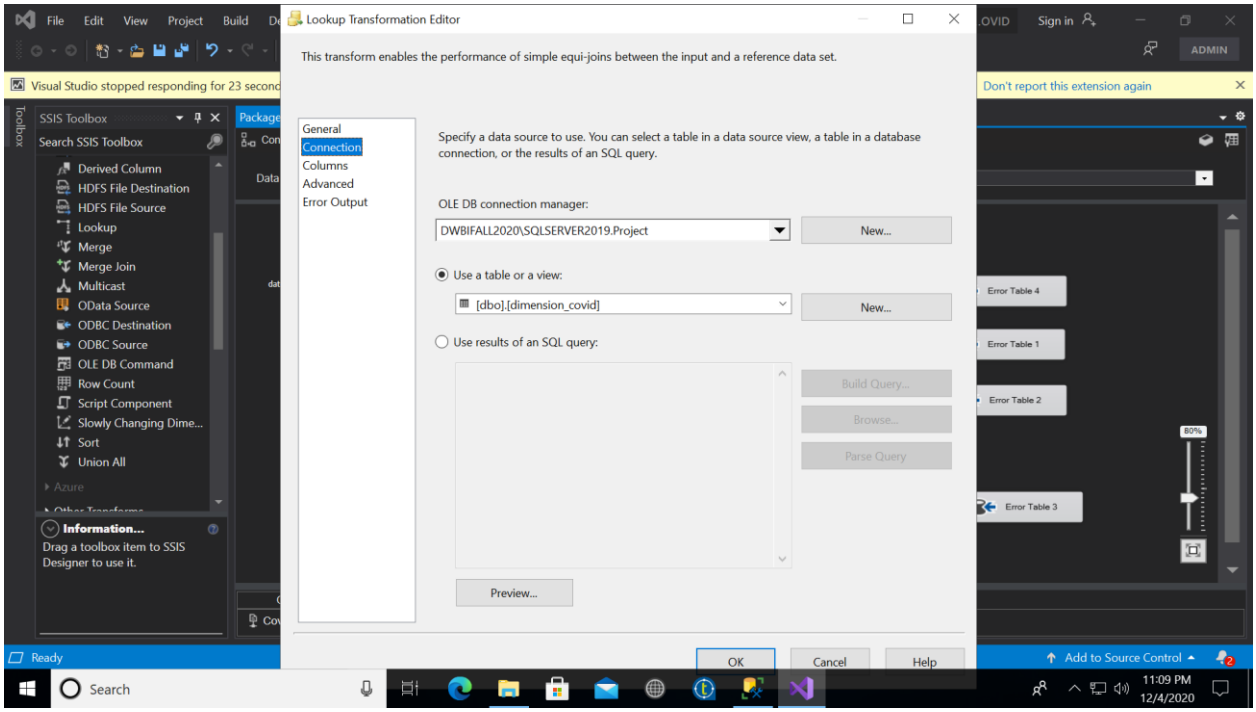


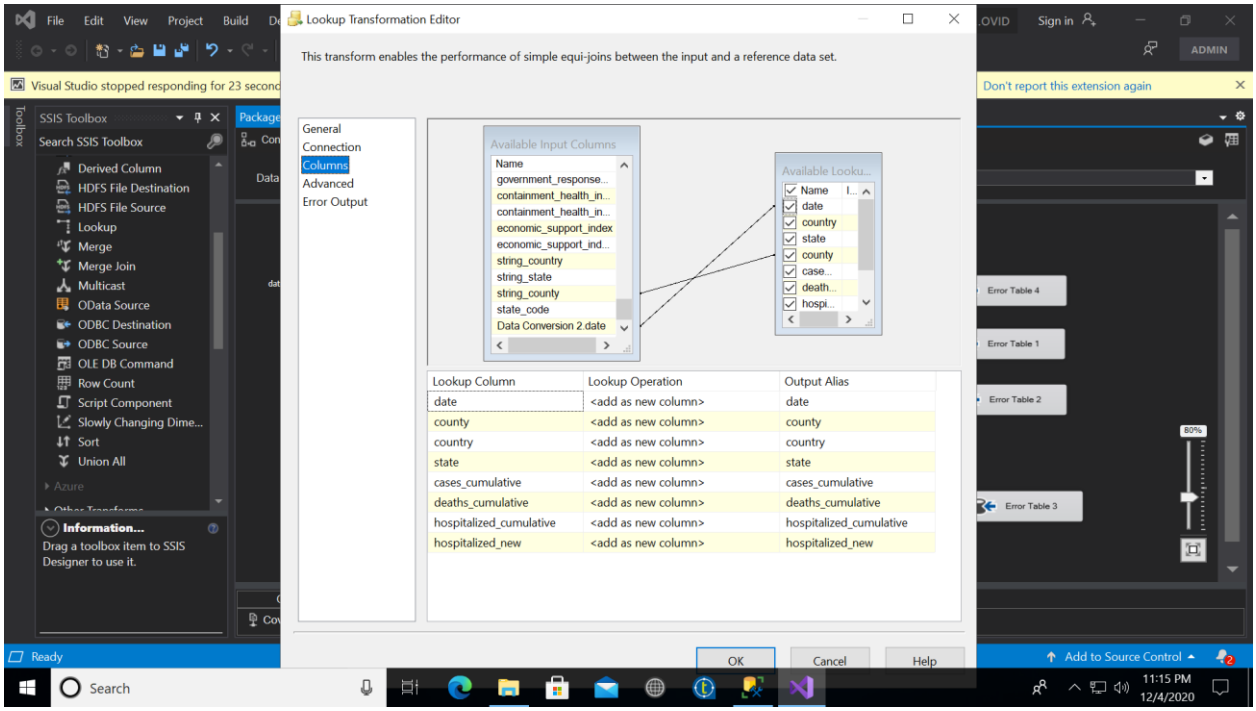
State code format check: If the state code has more than 2 characters then its going error table if not then going to dimension table





Duplicate check: duplicate entries are going to error table and the check is performed on ‘Date’ and ‘County’





Dimensional Model

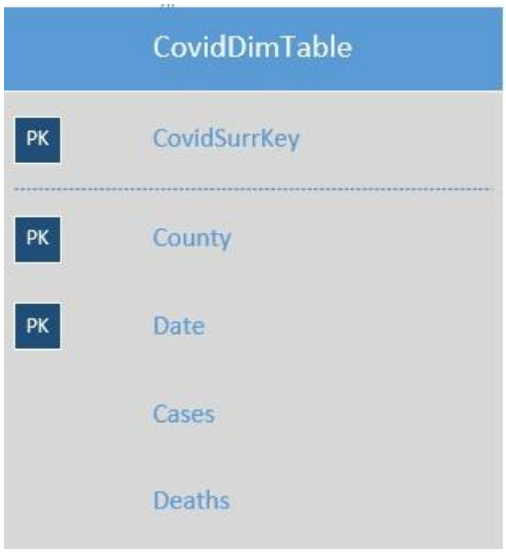
Our model is a Star Schema with 1 Fact Table, 3 Dimensional Tables, one for each data source and 1 Outrigger for region information.

The schemas for each of the Dimensional Tables are as follows

Weather

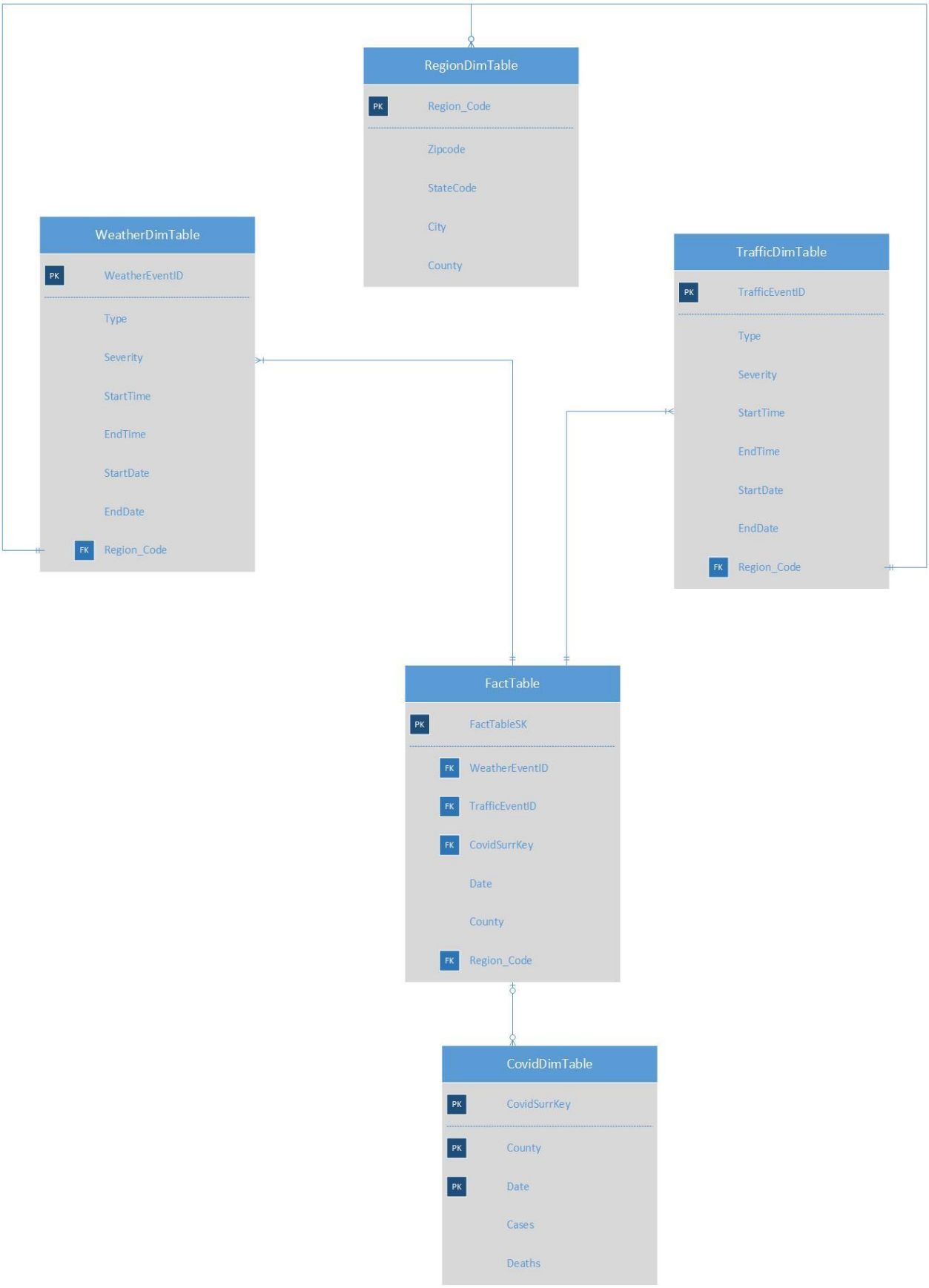
Traffic

Covid



Region Dim Table (Outrigger)

ER Diagram for Star Schema



Populating Fact Table

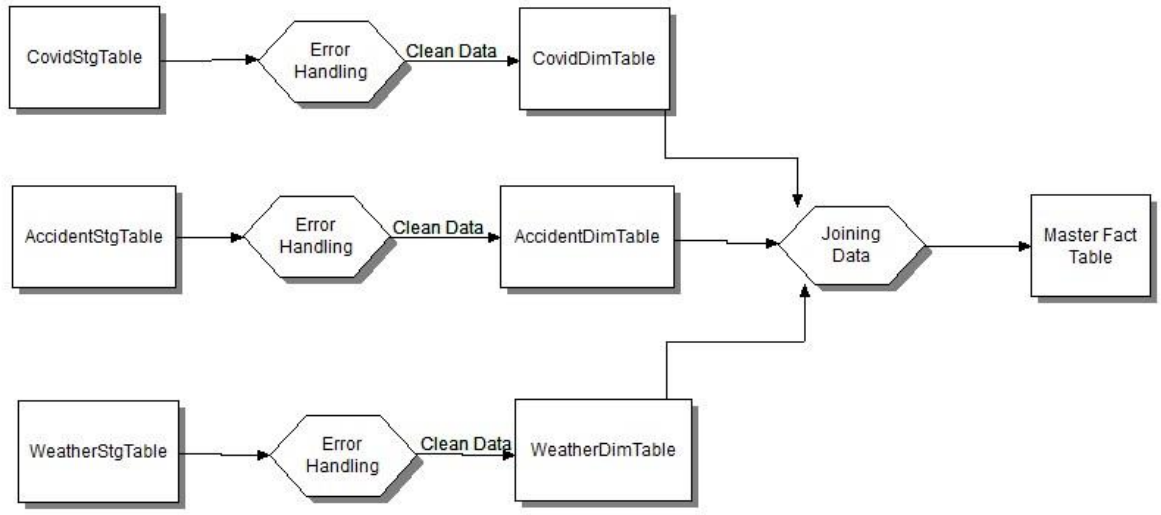
The fact table was populated by joining the weather and traffic data on region_code, and this will be joined to the covid data through county and date.

The Primary Key is a machine generated surrogate key, while the Alternate key is TrafficEventID+WeatherEventID + CovidSurrKey. Since Covid data only exists after 2020, the CovidSurrKey can be NULL.

There was considerable discussion whether there should be 2 fact tables – one before covid, consisting only of weather and traffic, and another after covid, consisting of some aggregations of weather and traffic, since Covid data is available only at a higher granularity (County) than either of those (Streets and Cities).

In the end, we felt it was better to have a single fact table as mentioned in the initial proposal to ensure better consistency of data. As a result, the covid data is a little redundant in the fact table as it would contain the same information in rows that have different cities but in the same county.

High Level View of population of the Fact Table (Initial Load)



SQL Script of the Join

```
insert into Fact_Table(EventId,Date,Region_Code,AccID,county,CovidID) (SELECT
K.EventId,K.StartDate AS 'Date',K.Region_Code,K.AccID,jc.county,jc.CovidID
FROM
(SELECT W.EventId,W.StartDate,W.Region_Code,A.EventId AS AccID FROM Dim_Weather  AS W
JOIN Dim_Accident AS A ON
W.Region_Code = A.Region_Code AND W.StartDate = A.StartDate) AS K
LEFT OUTER JOIN
(SELECT C.date,C.CovidID,R.Region_Code,R.County_Name AS
'County',C.cases_cumulative,C.deaths_cumulative FROM Dim_Covid AS C
JOIN Dim_Region AS R
ON C.county = R.County_Name) AS jc
ON
jc.date = K.StartDate AND jc.Region_Code = K.Region_Code)
```

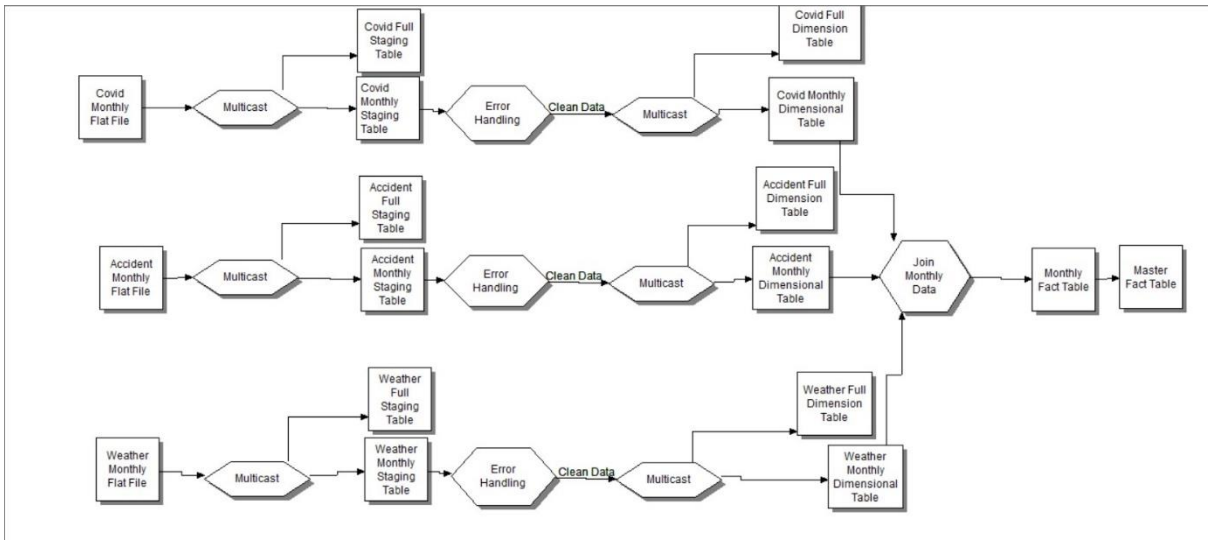
The above script was used for the initial load, to perform the join.

Future Loads

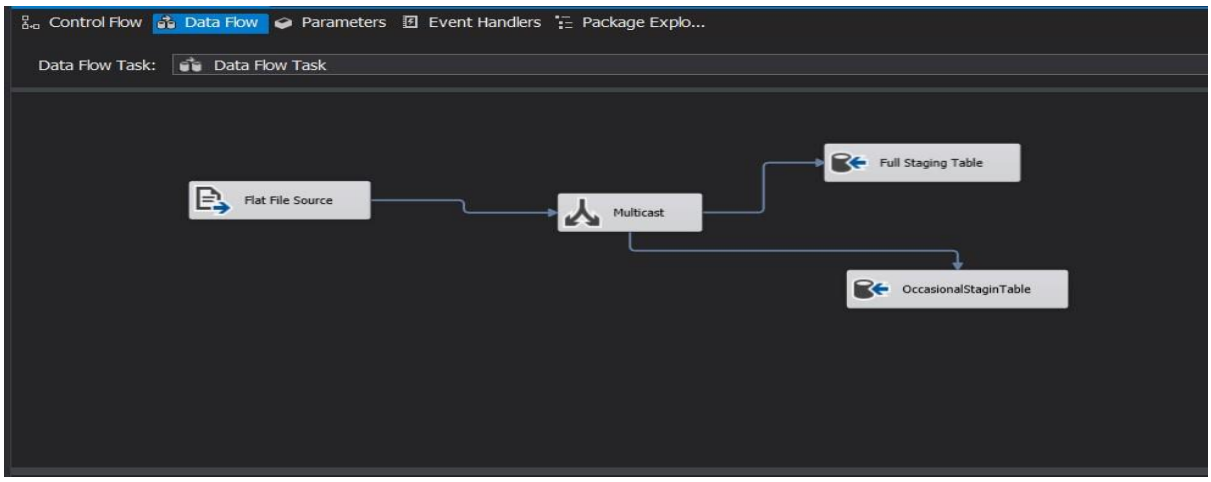
As can be seen, we implemented error handling on each of the dimensional table, which consequently ensures that only correct data will be going into the fact table.

For future loads, we created temporary staging and dimension tables for covid, weather and accidents, along with another temporary fact table. The reason to do this was to enable the staging tables, used for the initial load, to become an archive for the data that was loaded into it.

High Level view of Population of Fact Table (Monthly Loads)

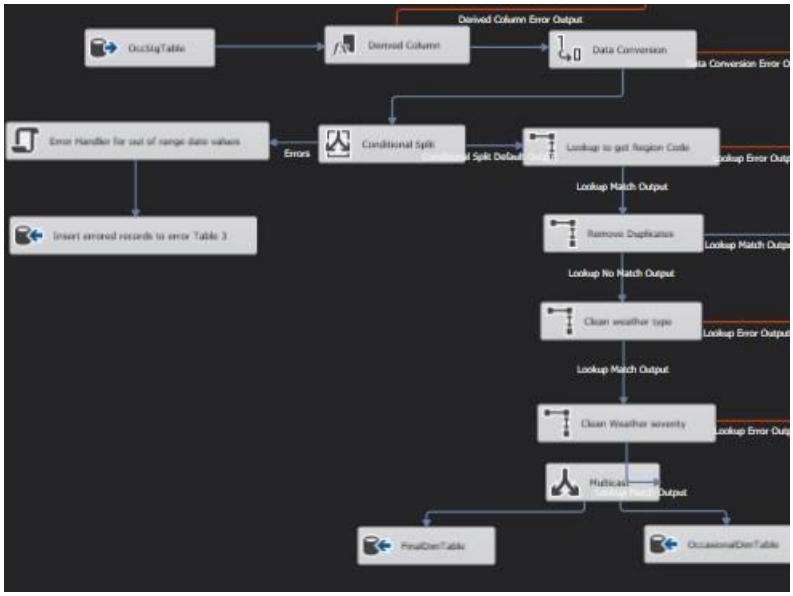


Monthly Flat file Multicast to Full Staging Table and Occasional Staging Table



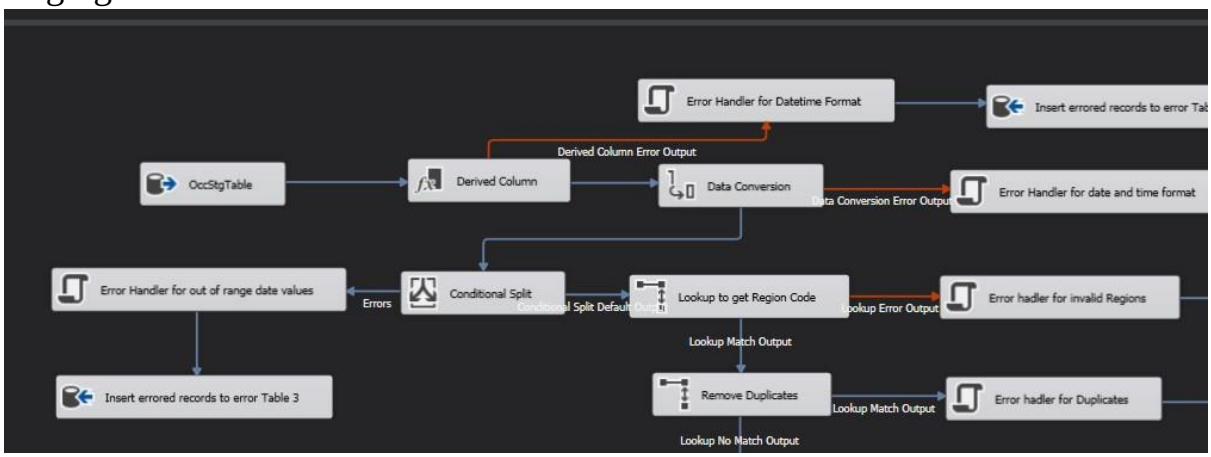
The multicast is done to ensure that we have a record of the monthly load in our archive, that is the full staging table. All the 3 data sources are handled similarly.

Error Handling for the Monthly loads

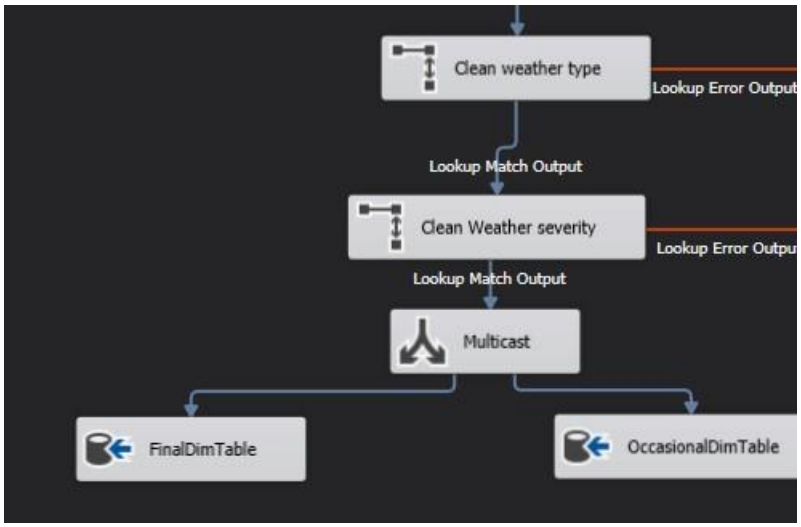


Once again, this is just the entire error handling we described previously that is just being used here, with 2 minor differences.

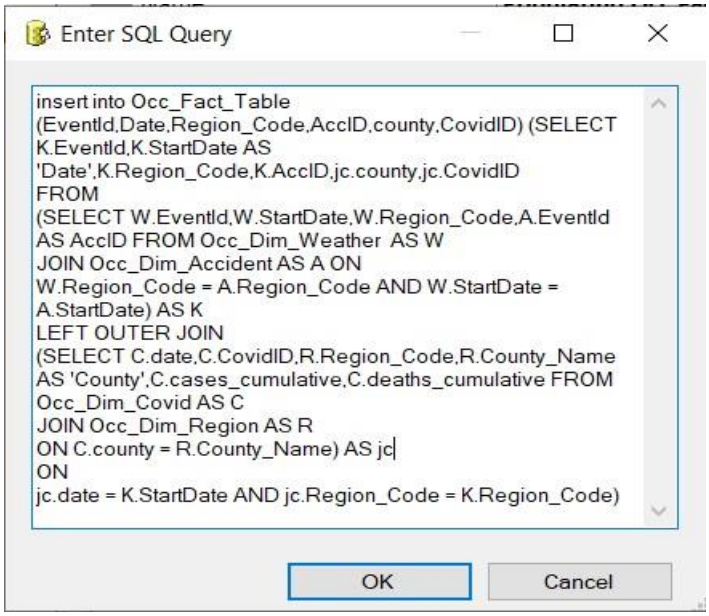
- The source table is the occasional staging table, instead of the FULL staging table.



- The valid rows after catching all errors, are multicasted to the Final Dimension Table and the Occasional Dimension Table

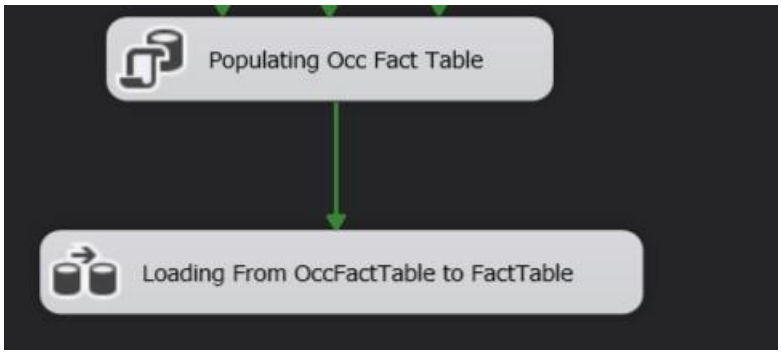


Populating the Occasional Fact Table



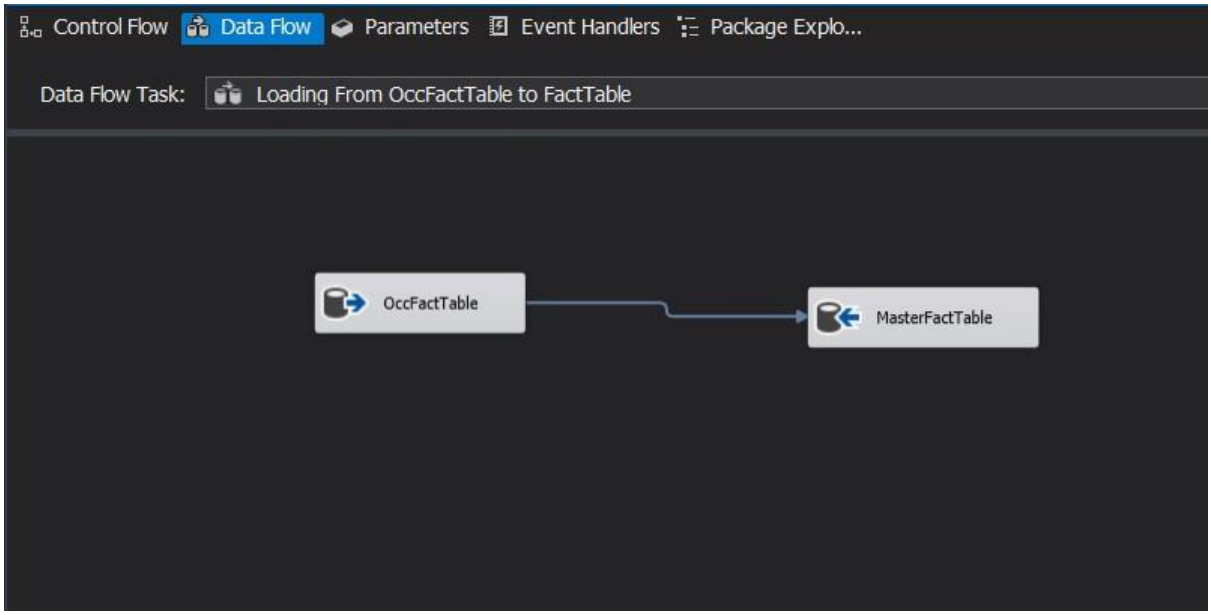
This is the SQL Script used to populate the occasional Fact Table. With this, we are ready to write the monthly data to the final fact table.

Control flow from the above Execute SQL Command to Data Flow from the Occasional Fact table to the Master Fact table



After obtaining the data from the 3 monthly Dimension tables and joining them, we load the data from the occasional Fact Table to the Master Fact Table

Data Flow



We took this approach as we felt it is important to maintain an archive of the data that we load into our master tables, in case we need it.

Ideally, the monthly tables are left as is until the next load, which would then be truncated and the same process is performed. This will facilitate any validation processes that can be performed on the most recently loaded/processed data, until the next load.

Data Marts

As mentioned in the proposal, we created the fact table to facilitate the aggregation of 2 important Data Marts - Pre-Covid and Post-Covid data.

Pre-Covid Data Mart

```
SELECT FactTableSK,EventId,Date,Region_Code,AccID INTO Pre_CovidDataMart FROM Fact_Table
WHERE CovidID IS NULL ORDER BY Date
```

With this we can create further views by joining the RegionDim Table and performing aggregations by geographic regions, all of which were done further in Tableau.

Post-Covid Data Mart

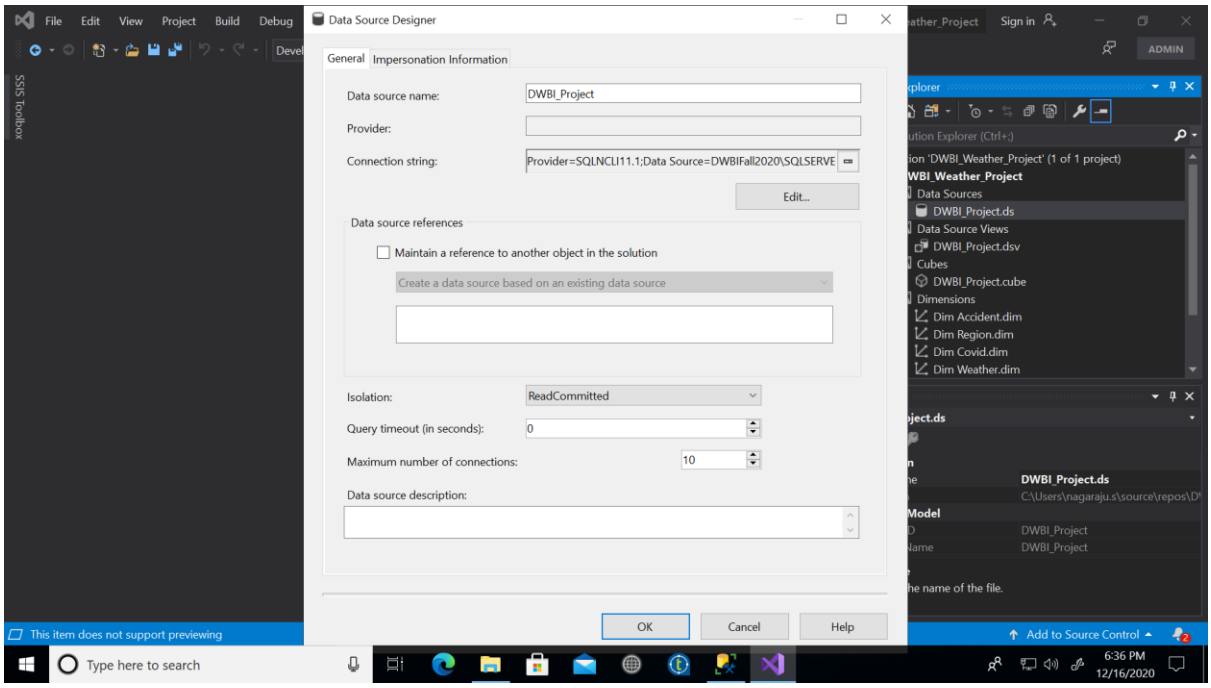
```
--PostCovid Data Mart
SELECT * INTO Post_CovidDataMart FROM Fact_Table WHERE CovidID IS NOT NULL ORDER BY Date
```

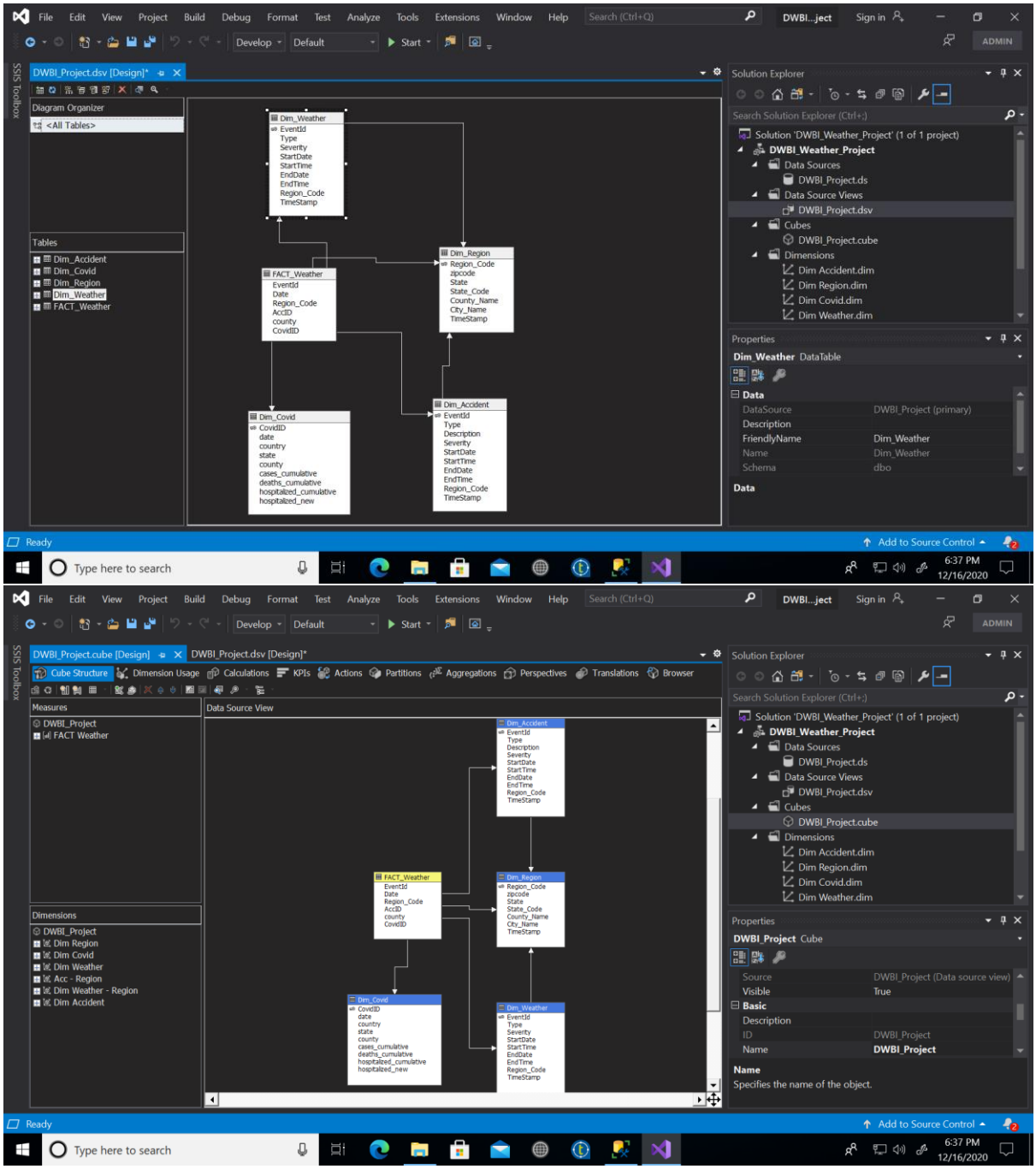
With the post Covid data, we can join the dimensional table, Weather and Accident, and use the Region tables to derive data and answer more specific questions.

We felt it is important to have a lot of options that let the users freely manipulate the data through joins, and answer any specific questions that they might have.

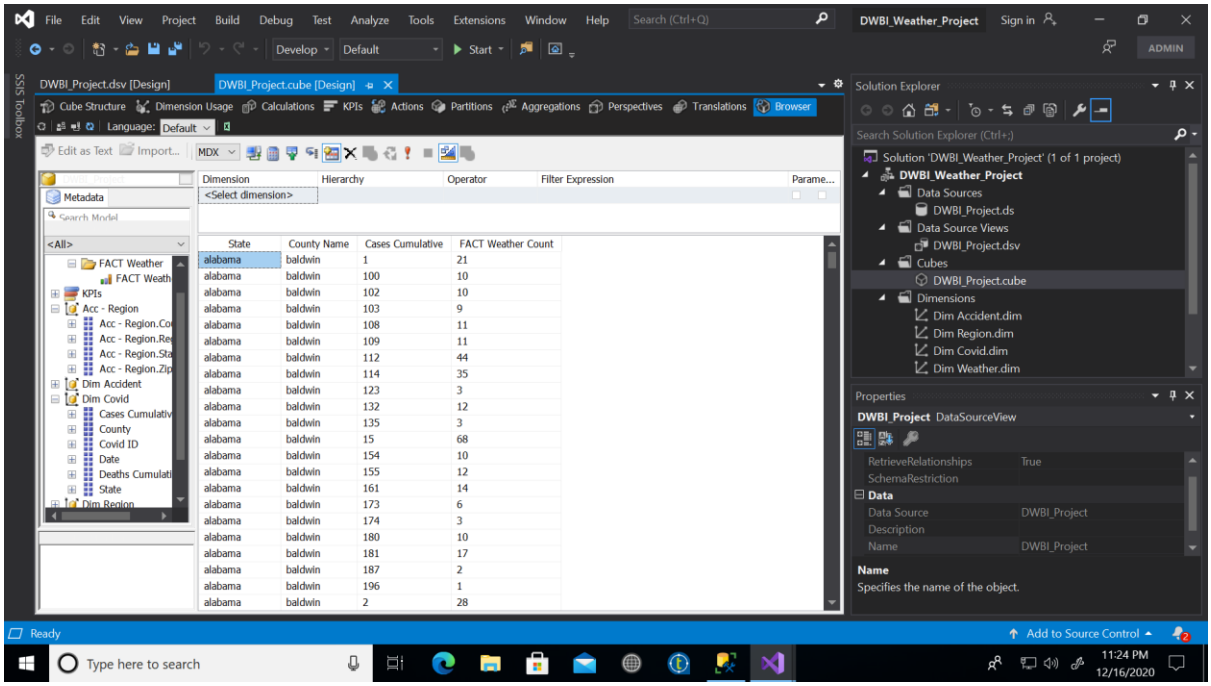
OLAP CUBE:

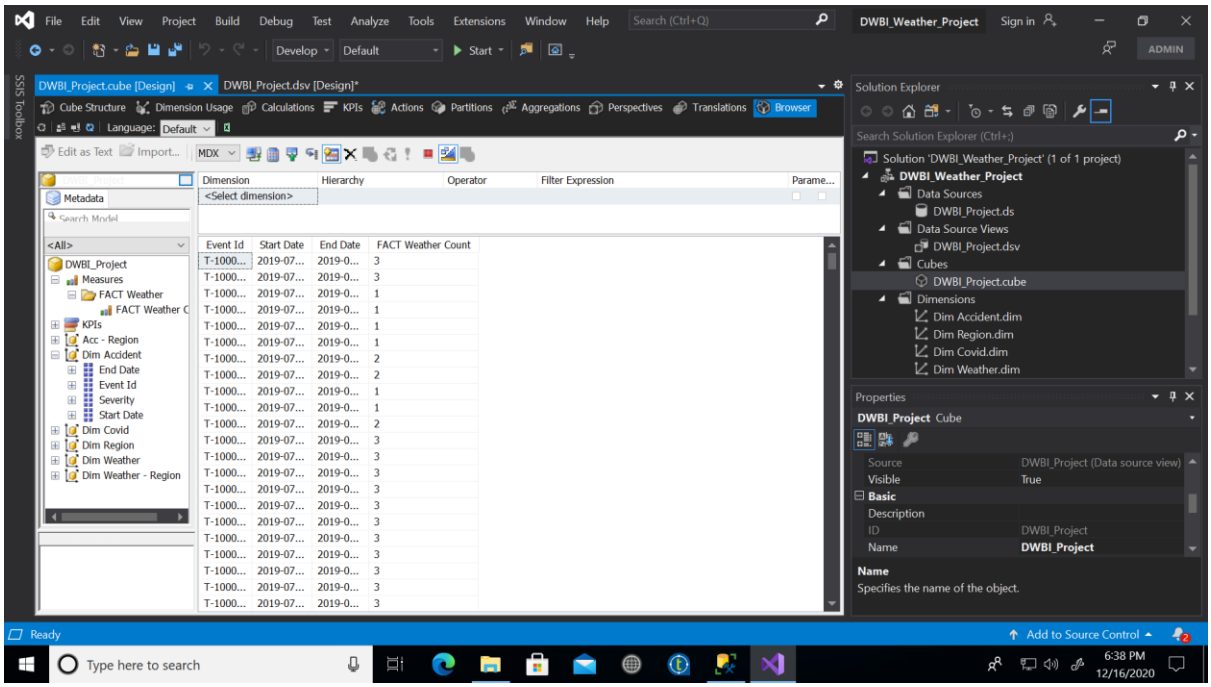
OLAP cube that would allow us to drill up and down and create various different distributions of data especially on different date parts (ex. Analysis of weather and traffic patterns on specific day of the year, day of the week, day of the month etc.)





In the below screenshot number of covid cases in each county is shown.


DWBI_Project.dsv [Design]
DWBI_Project.cube [Design]
Cube Structure
Dimension Usage
Calculations
KPIs
Actions
Partitions
Aggregations
Perspectives
Translations
Browser
Solution Explorer
DWBI_Weather_Project
Data Sources
DWBI_Project.ds
Data Source Views
DWBI_Project.dsv
Cubes
DWBI_Project.cube
Dimensions
Dim Accident.dim
Dim Region.dim
Dim Covid.dim
Dim Weather.dim
Properties
DWBI_Project DataSourceView
Ready
Add to Source Control
11:24 PM
12/16/2020

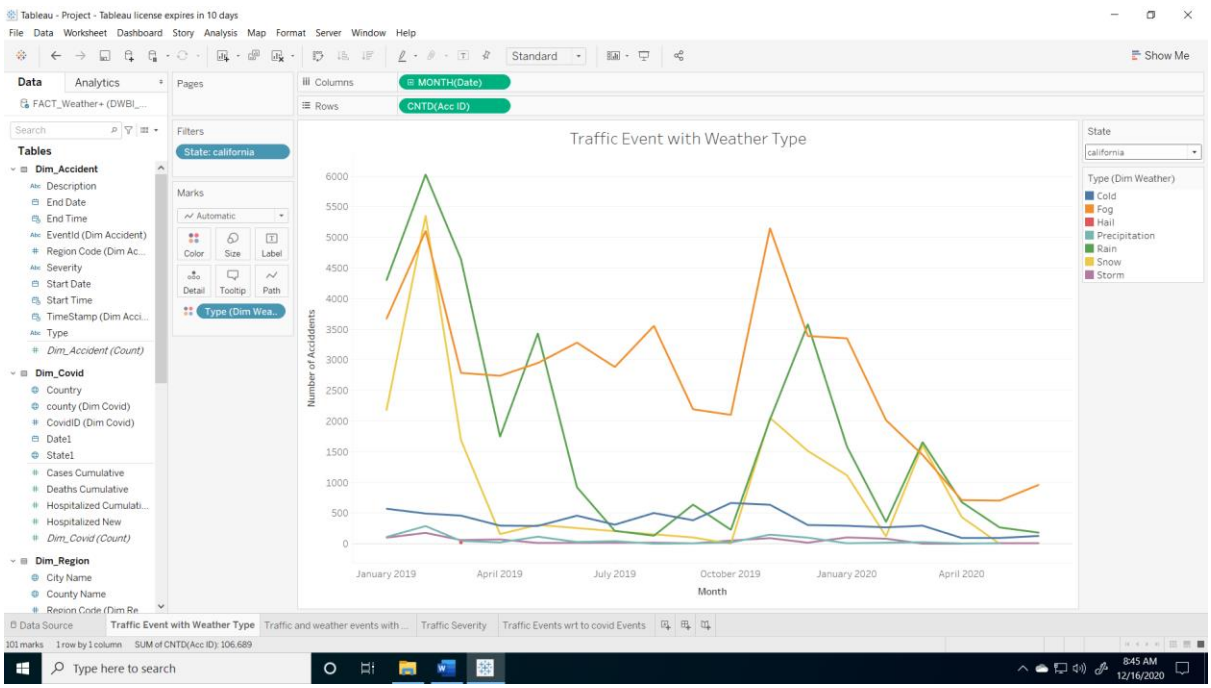

DWBI_Project.cube [Design]
DWBI_Project.dsv [Design]*
Cube Structure
Dimension Usage
Calculations
KPIs
Actions
Partitions
Aggregations
Perspectives
Translations
Browser
Solution Explorer
DWBI_Weather_Project
Data Sources
DWBI_Project.ds
Data Source Views
DWBI_Project.dsv
Cubes
DWBI_Project.cube
Dimensions
Dim Accident.dim
Dim Region.dim
Dim Covid.dim
Dim Weather.dim
Properties
DWBI_Project Cube
Ready
Add to Source Control
6:38 PM
12/16/2020

Visualization

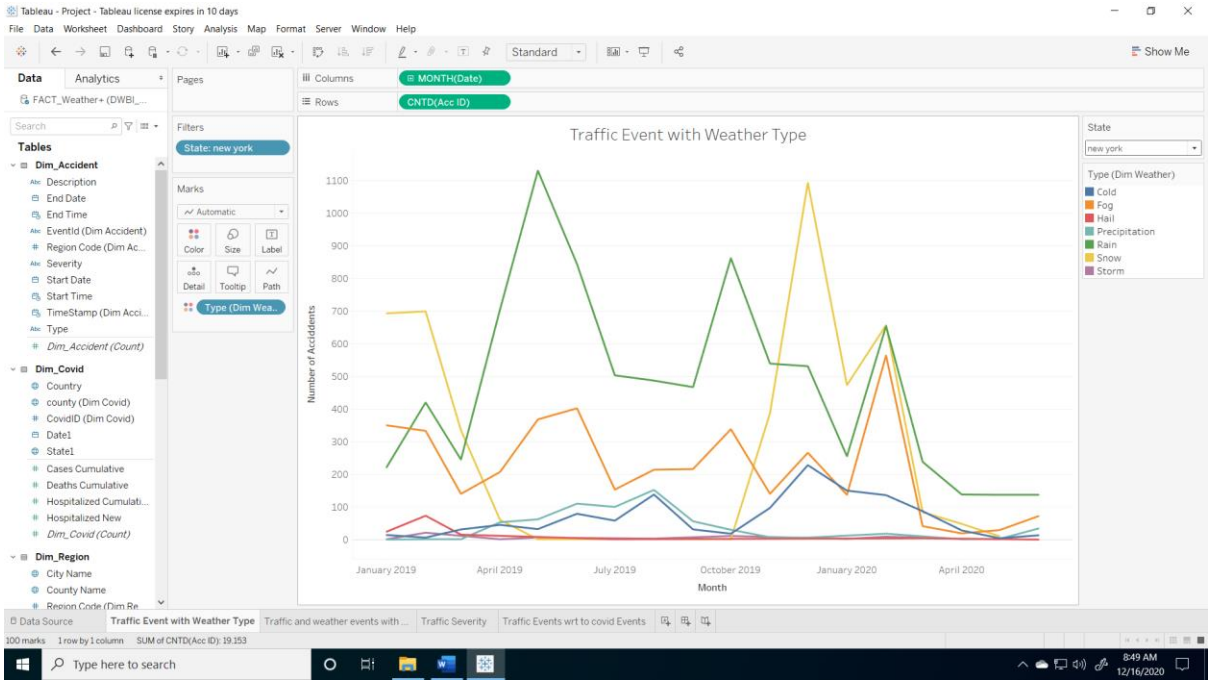
Using Tableau, we created some powerful visualizations from the Data Marts that we created. For example, we will be creating a Dashboard that will help us understand how traffic patterns changed with increasing covid cases in a particular area. Also, how weather patterns might have influenced the progression of Covid cases by county, and how these 3 correlates. As we keep building such visualizations, there are bound to be some more meaningful ideas that come to us which we plan to include.

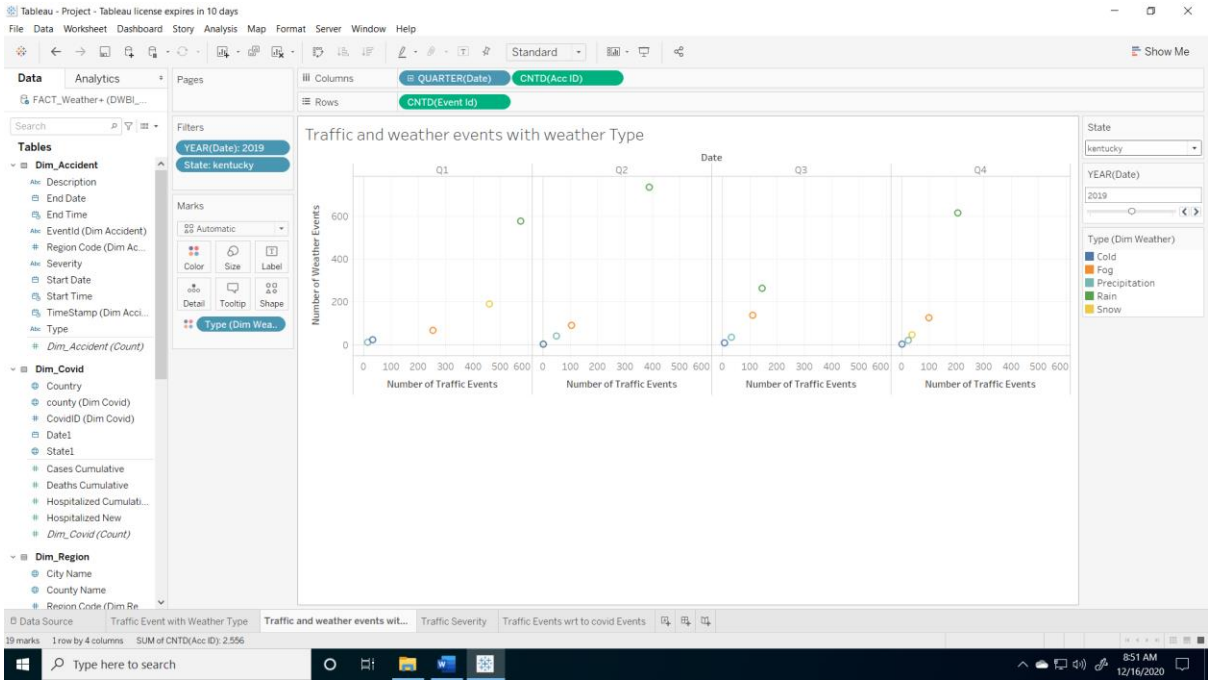
Number of Traffic events based on Weather Types:

The Below plot shows the traffic events that took place in California from 2019 to 2020 June. As we can see there's a relationship between the climatic condition and Traffic events. The weather conditions play a vital role in the traffic condition.



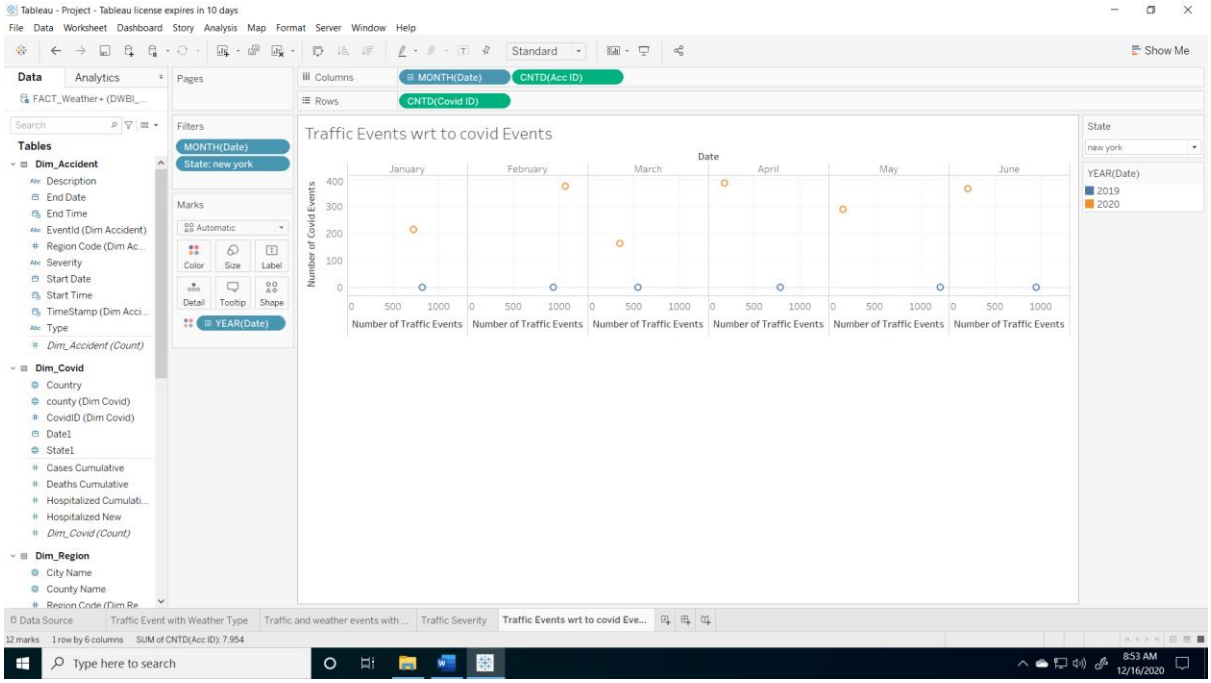
This a same plot for the state of New York. Here as well we can observe the role played by weather conditions in the traffic events.





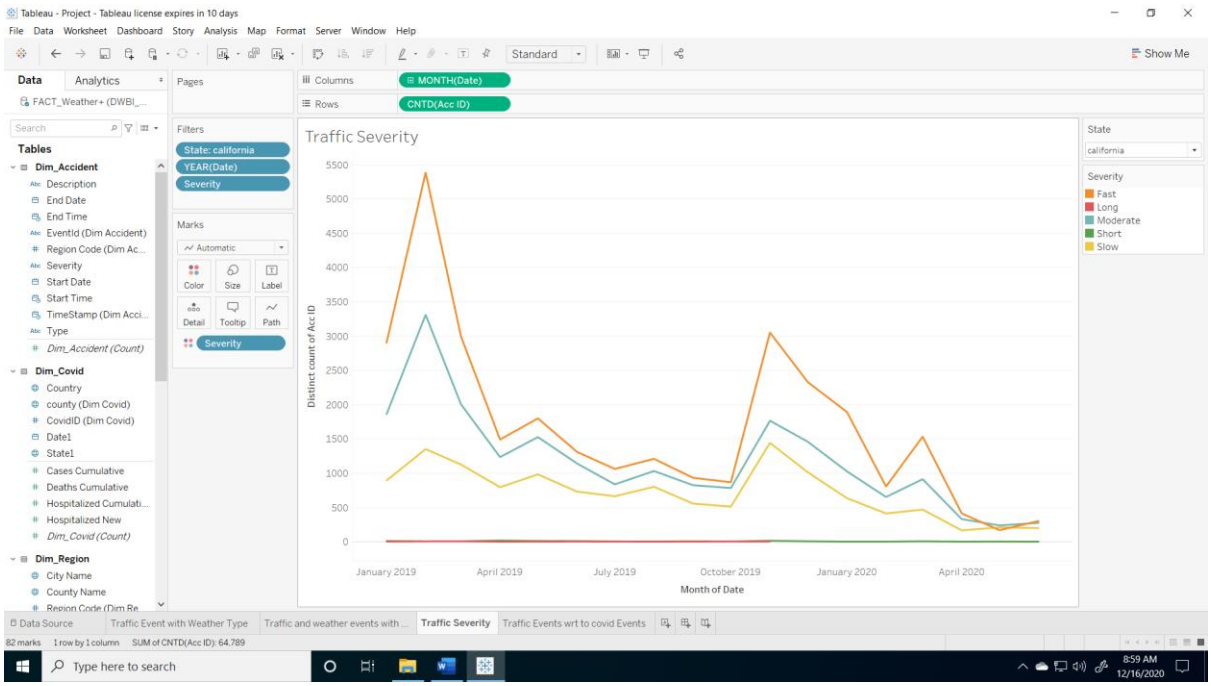
The Above plot shows the quarter wise accident and weather events that took place in the state of Kentucky. Based on the observation is clear to say rain is the major reason for accidents to occur in majority of states.

Traffic Events with respect to COVID-19



The plot shows the number of traffic incidents for the months of Jan-June for 2019 and 2020.

Covid-19 played a major role in traffic accidents that took place in 2020, as we can see the number of covid events started to increase from the month of March, so the overall traffic accidents can be seen to have reduced. Since much of the population was under Lockdown.



Revision History

Changes	Versions	Performed by
Initial Proposal	1	GRP 9
• Added more detailed Covid data- now contains demographic information. • Column wise error handling. • Transformation of date-time into date and time fields. • Splitting of data to accommodate Monthly loads.	2	GRP 9
• Completed Staging Tables, Dimensional Tables and Error handling • Reduced the effective date from 2016 to 2019 for weather and covid • Cleaner integration of Covid data into the Fact Table • Slight changes to the ER Diagram, with the addition of RegionDimTable as Outrigger	3	GRP 9
• Population of fact table • Creation of monthly staging, dimensional and fact tables to implement future loads • Creation of Data Marts, OLAP Cube • Created Data Visualizations on Tableau	4	GRP 9

Appendix

Initial Draft Notes

==========

You will need to make sure you make provisions for adding data this may require splitting the files etc.

Try to find an additional covid data set just so the second mart has a little more info (would be nice)

You will need to create some bad data to test error handling

Good start

Version 2

Shouldn’t the covid data be part of the fact. I don’t see what the facts are in the picture as it stands now

Version 3

You are on the right track keep progressing.

92.5