

DAMG 7370 Group 10 Final Project

Part 1 :

Ydata profiling

Motor Vehicle Collisions - Crashes New York Dataset: New York

Dataset Size: The dataset contains 2075427 entries (rows) and 29 columns.

	CRASH DATE	CRASH TIME	BOROUGH	ZIP CODE	LATITUDE	LONGITUDE	LOCATION	ON STREET NAME	CROSS STREET NAME	OFF STREET NAME	...	CONTRIBUTING FACTOR VEHICLE 2	CONTRIBUTING FACTOR VEHICLE 3
0	09/11/2021	2:39	NaN	NaN	NaN	NaN	NaN	WHITESTONE EXPRESSWAY	20 AVENUE	NaN	...	Unspecified	NaN
1	03/26/2022	11:45	NaN	NaN	NaN	NaN	NaN	QUEENSBORO BRIDGE UPPER	NaN	NaN	...	NaN	NaN
2	06/29/2022	6:55	NaN	NaN	NaN	NaN	NaN	THROGS NECK BRIDGE	NaN	NaN	...	Unspecified	NaN
3	09/11/2021	9:35	BROOKLYN	11208.0	40.667202	-73.866500	(40.667202, -73.8665)	NaN	NaN	1211 LORING AVENUE	...	NaN	NaN
4	12/14/2021	8:13	BROOKLYN	11233.0	40.683304	-73.917274	(40.683304, -73.917274)	SARATOGA AVENUE	DECATUR STREET	NaN	...	NaN	NaN

Column Types:

- There are 18 columns of type Text, which are likely categorical.
- There are 11 columns of type integer, which are numerical with decimal values.

Missing Values:

- BOROUGH has 645,746 missing values, while ZIP CODE has 645,996 missing values.
- LATITUDE, LONGITUDE, and LOCATION each exhibit 233,626 missing values.
- ON STREET NAME is missing in 440,569 instances, and CROSS STREET NAME is missing in 784,436 instances.
- OFF STREET NAME has a significant 1,727,231 missing values.
- NUMBER OF PERSONS INJURED is absent in 18 cases, and NUMBER OF PERSONS KILLED is absent in 31 cases.
- CONTRIBUTING FACTOR VEHICLE 1 has 6,802 missing values, whereas CONTRIBUTING FACTOR VEHICLE 2 has 321,736 missing values.
- CONTRIBUTING FACTOR VEHICLE 3, CONTRIBUTING FACTOR VEHICLE 4, and CONTRIBUTING FACTOR VEHICLE 5 each lack over 1.9 million entries.
- VEHICLE TYPE CODE 1 is missing in 13,691 instances, and VEHICLE TYPE CODE 2 is missing in 396,691 instances.
- VEHICLE TYPE CODE 3, VEHICLE TYPE CODE 4, and VEHICLE TYPE CODE 5 each have over 1.9 million missing values.
- COLLISION_ID stands out with no missing values.

The Source Column names for the New York Dataset is as follows:

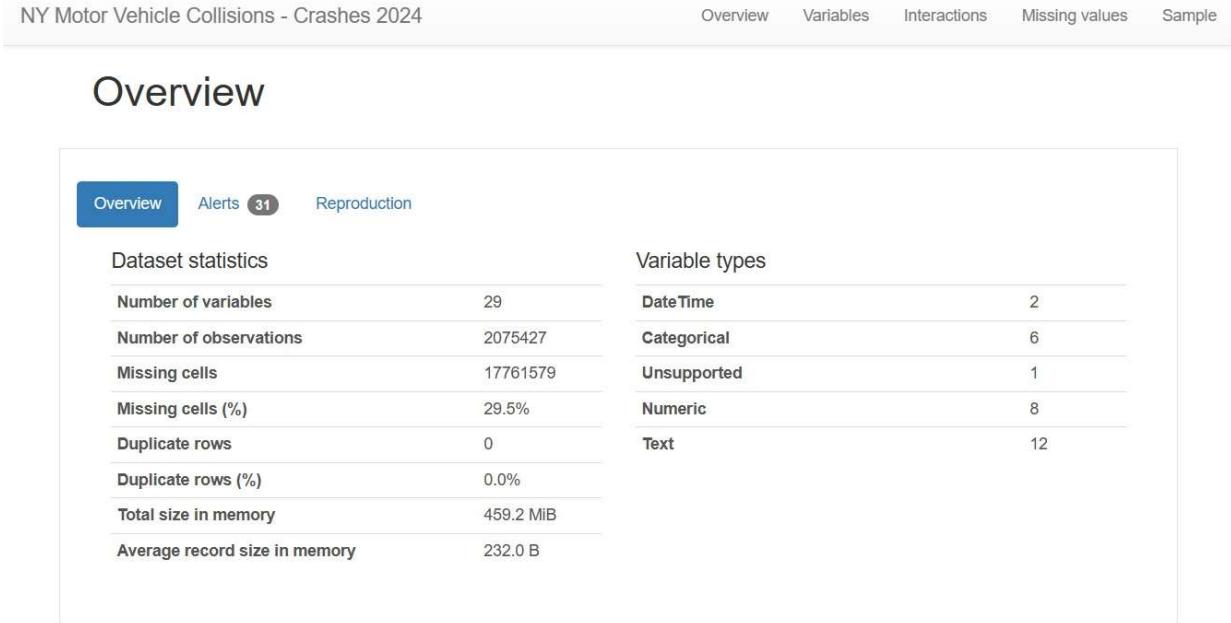


df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075427 entries, 0 to 2075426
Data columns (total 29 columns):
 #   Column           Dtype  
--- 
 0   CRASH DATE      object  
 1   CRASH TIME      object  
 2   BOROUGH         object  
 3   ZIP CODE        object  
 4   LATITUDE        float64 
 5   LONGITUDE       float64 
 6   LOCATION         object  
 7   ON STREET NAME  object  
 8   CROSS STREET NAME object  
 9   OFF STREET NAME object  
 10  NUMBER OF PERSONS INJURED float64 
 11  NUMBER OF PERSONS KILLED float64 
 12  NUMBER OF PEDESTRIANS INJURED int64  
 13  NUMBER OF PEDESTRIANS KILLED int64  
 14  NUMBER OF CYCLIST INJURED int64  
 15  NUMBER OF CYCLIST KILLED int64  
 16  NUMBER OF MOTORIST INJURED int64  
 17  NUMBER OF MOTORIST KILLED int64  
 18  CONTRIBUTING FACTOR VEHICLE 1 object  
 19  CONTRIBUTING FACTOR VEHICLE 2 object  
 20  CONTRIBUTING FACTOR VEHICLE 3 object  
 21  CONTRIBUTING FACTOR VEHICLE 4 object  
 22  CONTRIBUTING FACTOR VEHICLE 5 object  
 23  COLLISION_ID    int64  
 24  VEHICLE TYPE CODE 1 object  
 25  VEHICLE TYPE CODE 2 object  
 26  VEHICLE TYPE CODE 3 object  
 27  VEHICLE TYPE CODE 4 object  
 28  VEHICLE TYPE CODE 5 object  
dtypes: float64(4), int64(7), object(18)
memory usage: 459.2+ MB
```

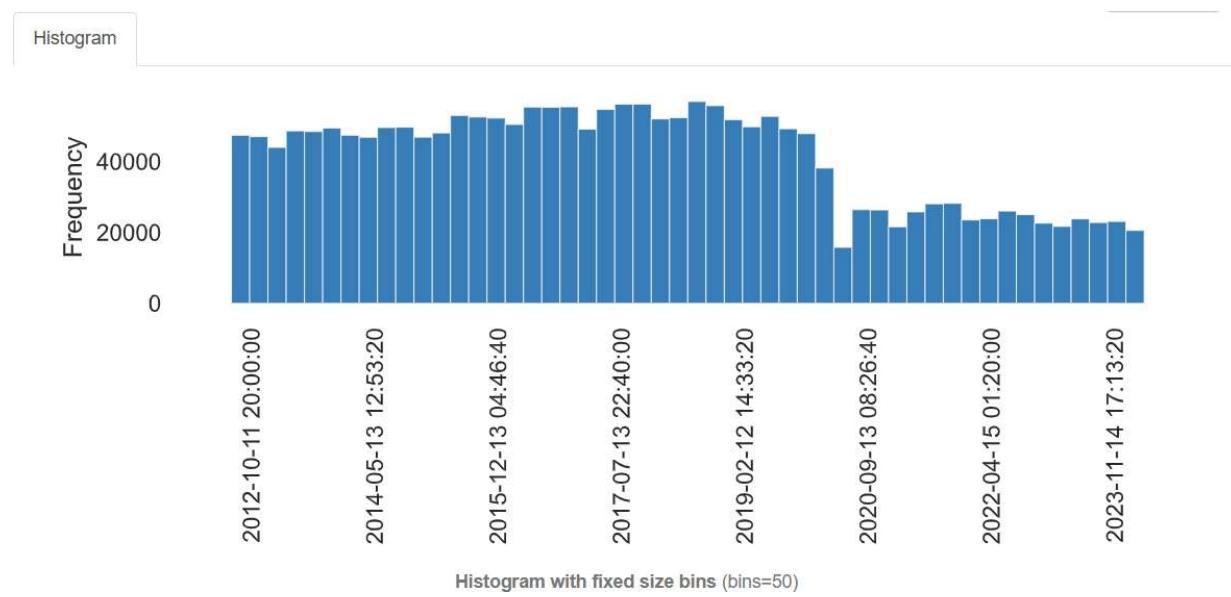
[2075427 rows x 29 columns]

The following is the overview of the whole dataset which consists of all the missing cells, rows, data types,



etc.

For example



The above histogram represents the number of crashes which took place between year 2012 - year 2023

Traffic Crashes - Crashes Chicago

Dataset: Chicago

Dataset Size: The dataset contains 817,723 entries (rows) and 48 columns.

	CRASH_RECORD_ID	CRASH_DATE_EST_I	CRASH_DATE	POSTED_SPEED_LIMIT	TRAFFIC_CONTROL_DEVICE	DEVICE_CONDITION	WEATH
0	6c1659069e9c6285a650e70d6f9b574ed5f64c12888479...	NaN	08/18/2023 12:50:00 PM	15	OTHER	FUNCTIONING PROPERLY	
1	5f54a59fc087b12ae5b1acff96a3caf4f2d37e79f8db4...	NaN	07/29/2023 02:45:00 PM	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	
2	61fcb8c1eb522a6469b460e2134df3d15f82e81fd93e9c...	NaN	08/18/2023 05:58:00 PM	30	NO CONTROLS	NO CONTROLS	
3	004cd14d0303a9163aad69a2d7f341b7da2a8572b2ab33...	NaN	11/26/2019 08:38:00 AM	25	NO CONTROLS	NO CONTROLS	
4	a1d5f0ea90897745365a4cbb06cc60329a120d89753fac...	NaN	08/18/2023 10:45:00 AM	20	NO CONTROLS	NO CONTROLS	

5 rows x 48 columns

Column Types:

- There are 31 columns of type object, which are likely categorical or textual in nature.
- There are 11 columns of type float64, which are likely numerical with decimal values.
- There are 6 columns of type int64, which are likely numerical with integer values.

Missing Values:

- Several columns have a significant number of missing values.
- For example: CRASH_DATE_EST_I has 756,594 missing values, which is about 92.5% of the total entries.
- LANE_CNT has 618,714 missing values, which is about 75.7% of the total entries.
- INTERSECTION RELATED_I has 630,174 missing values, which is about 77.1% of the total entries.
- Columns like STREET_DIRECTION, STREET_NAME, BEAT_OF_OCCURRENCE, and MOST_SEVERE_INJURY have relatively few missing values.
- Some columns like PHOTOS_TAKEN_I, STATEMENTS_TAKEN_I, DOORING_I, WORK_ZONE_I, WORK_ZONE_TYPE, and WORKERS_PRESENT_I have a very high percentage of missing values, which might indicate that these events are rare or not commonly reported.

The Source Column names for the New York Dataset is as follows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 817723 entries, 0 to 817722
Data columns (total 48 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   CRASH_RECORD_ID    817723 non-null   object 
 1   CRASH_DATE_EST_I   61129 non-null   object 
 2   CRASH_DATE         817723 non-null   object 
 3   POSTED_SPEED_LIMIT 817723 non-null   int64  
 4   TRAFFIC_CONTROL_DEVICE 817723 non-null   object 
 5   DEVICE_CONDITION    817723 non-null   object 
 6   WEATHER_CONDITION   817723 non-null   object 
 7   LIGHTING_CONDITION  817723 non-null   object 
 8   FIRST_CRASH_TYPE   817723 non-null   object 
 9   TRAFFICWAY_TYPE    817723 non-null   object 
 10  LANE_CNT           199009 non-null   float64
 11  ALIGNMENT          817723 non-null   object 
 12  ROADWAY_SURFACE_COND 817723 non-null   object 
 13  ROAD_DEFECT        817723 non-null   object 
 14  REPORT_TYPE        793409 non-null   object 
 15  CRASH_TYPE         817723 non-null   object 
 16  INTERSECTION_RELATED_I 187549 non-null   object 
 17  NOT_RIGHT_OF_WAY_I  37708 non-null   object 
 18  HIT_AND_RUN_I      255949 non-null   object 
 19  DAMAGE              817723 non-null   object 
 20  DATE_POLICE_NOTIFIED 817723 non-null   object 
 21  PRIM_CONTRIBUTORY_CAUSE 817723 non-null   object 
 22  SEC_CONTRIBUTORY_CAUSE 817723 non-null   object 
 23  STREET_NO           817723 non-null   int64  
 24  STREET_DIRECTION    817719 non-null   object 
 25  STREET_NAME         817722 non-null   object 
 26  BEAT_OF_OCCURRENCE  817718 non-null   float64
 27  PHOTOS_TAKEN_I     10775 non-null   object 
 28  STATEMENTS_TAKEN_I  18258 non-null   object 
 29  DOORING_I           2512 non-null   object 
 30  WORK_ZONE_I         4670 non-null   object
```

```
31 WORK_ZONE_TYPE          3618 non-null    object
32 WORKERS_PRESENT_I       1194 non-null    object
33 NUM_UNITS                817723 non-null   int64
34 MOST_SEVERE_INJURY      815931 non-null   object
35 INJURIES_TOTAL          815943 non-null   float64
36 INJURIES_FATAL          815943 non-null   float64
37 INJURIES_INCAPACITATING 815943 non-null   float64
38 INJURIES_NON_INCAPACITATING 815943 non-null   float64
39 INJURIES_REPORTED_NOT_EVIDENT 815943 non-null   float64
40 INJURIES_NO_INDICATION 815943 non-null   float64
41 INJURIES_UNKNOWN         815943 non-null   float64
42 CRASH_HOUR               817723 non-null   int64
43 CRASH_DAY_OF_WEEK        817723 non-null   int64
44 CRASH_MONTH              817723 non-null   int64
45 LATITUDE                  812108 non-null   float64
46 LONGITUDE                 812108 non-null   float64
47 LOCATION                  812108 non-null   object
dtypes: float64(11), int64(6), object(31)
memory usage: 299.5+ MB
```

[817723 rows x 48 columns]

The following is the overview of the whole dataset which consists of all the missing cells, rows, data types, etc.

Chicago Traffic Crashes - Crashes 2024

Overview Variables Interactions Missing values Sample



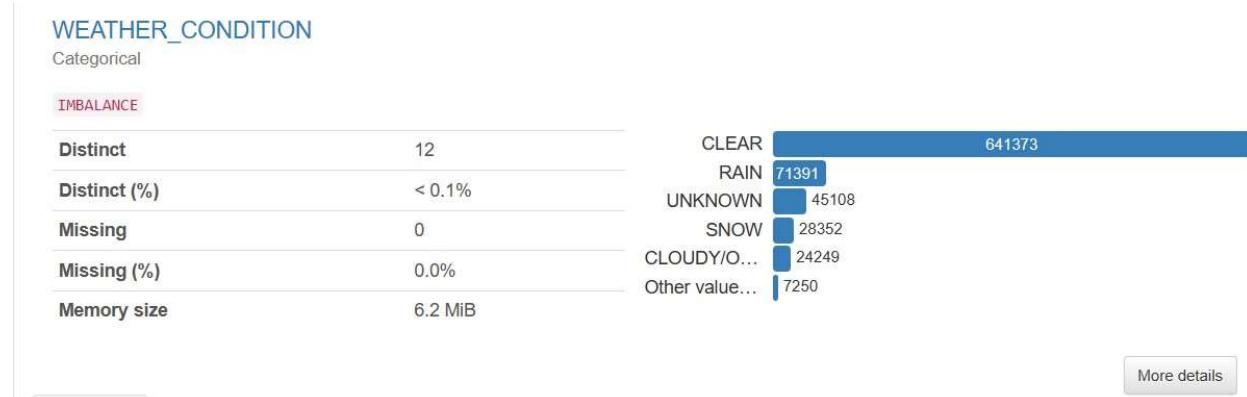
Overview

Overview	Alerts 35	Reproduction
Dataset statistics		Variable types
Number of variables		Text
Number of observations		3
Missing cells		Boolean
Missing cells (%)		2
Duplicate rows		DateTime
Duplicate rows (%)		Numeric
Total size in memory		Categorical
Average record size in memory		15
		19

Shruti Randive
NUID 002740632

For Example:

If we want the details of the weather condition column we can get it through the following data



So we can see the max length, min length etc from the below data along with the sample rows

Overview	Categories	Words	Characters	Length	Characters and Unicode	Unique	Sample
				Max length	24	Unique	1st row
				Median length	5	0	CLEAR
				Mean length	5.3462053	Unique (%)	2nd row
				Min length	4	0.0%	CLEAR
					Total characters 4371715		3rd row
					Distinct characters 25		4th row
					Distinct categories 3	?	5th row
					Distinct scripts 2	?	
					Distinct blocks 1	?	
					The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.		

Austin Crash Report Data - Crash Level Records

Dataset: Austin

Dataset Size: The dataset contains 147750 entries (rows) and 54 columns.

	crash_id	crash_fatal_fl	crash_date	crash_time	case_id	rpt_latitude	rpt_longitude	rpt_block_num	rpt_street_pfx	rpt_street_name	...	pedestrian_serious_injury_c
0	13762420	N	03/30/2014 AM	10:58:00	140890874	NaN	NaN	NaN	NaN	3707 MANCHACA	...	
1	13777334	N	03/27/2014 PM	01:07:00	140860852	NaN	NaN	3400	NaN	PALM WAY TO MOPAC NB RAMP	...	
2	13777441	N	03/28/2014 PM	03:42:00	140871196	NaN	NaN	8704	NaN	BALCONES CLUB DR	...	
3	13797332	N	04/09/2014 PM	02:09:00	140991015	NaN	NaN	8000	NaN	E US 290 HWY SVRD EB	...	
4	13795604	N	04/07/2014 PM	06:00:00	140971248	NaN	NaN	200	W	BEN WHITE	...	

5 rows × 54 columns

Column Types:

There are 23 columns of type object, which are categorical or textual in nature. There are 14 columns of type float64, which are numerical with decimal values. There are 17 columns of type int64, which are numerical with integer values.

Missing Values:

- Several columns have a significant number of missing values. For example: rpt_latitude has 137,456 missing values, which is about 93.0% of the total entries.
- rpt_longitude has 137,456 missing values, which is about 93.0% of the total entries.
- street_nbr has 87,038 missing values, which is about 58.9% of the total entries.
- street_name_2 has 81,474 missing values, which is about 55.1% of the total entries.
- micromobility_fl has 147,439 missing values, which is about 99.8% of the total entries.
- Columns like crash_id, crash_fatal_fl, crash_date, and crash_time have no missing values.
- Some columns like case_id, rpt_block_num, and street_name have a moderate number of missing values.
- Overall, the dataset contains a varied range of missing values across different columns, indicating potential data quality issues and areas for further investigation.

The following are the source columns of the dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 147750 entries, 0 to 147749
Data columns (total 54 columns):
 #   Column           Non-Null Count  Dtype  
 ---  -- 
 0   crash_id         147750 non-null   int64  
 1   crash_fatal_flg  147750 non-null   object  
 2   crash_date       147750 non-null   object  
 3   crash_time       147750 non-null   object  
 4   case_id          145892 non-null   object  
 5   rpt_latitude     10294 non-null    float64 
 6   rpt_longitude    10294 non-null    float64 
 7   rpt_block_num    128139 non-null   object  
 8   rpt_street_pfx   79945 non-null   object  
 9   rpt_street_name  147747 non-null   object  
 10  rpt_street_sfx  97410 non-null   object  
 11  crash_speed_limit 147748 non-null   float64 
 12  road_constr_zone_flg 147748 non-null   object  
 13  latitude         145507 non-null   float64 
 14  longitude        145507 non-null   float64 
 15  street_name      147748 non-null   object  
 16  street_nbr       60712 non-null   float64 
 17  street_name_2    66276 non-null   object  
 18  street_nbr_2     0 non-null      float64 
 19  crash_sev_id     147750 non-null   int64  
 20  sus_serious_injry_cnt 147750 non-null   int64  
 21  nonincap_injry_cnt 147749 non-null   float64 
 22  poss_injry_cnt   147749 non-null   float64 
 23  non_injry_cnt    147749 non-null   float64 
 24  unkn_injry_cnt   147748 non-null   float64 
 25  tot_injry_cnt    147748 non-null   float64 
 26  death_cnt        147750 non-null   int64  
 27  accident_fatal_flg 147747 non-null   float64
```

```
27 contrib_factr_pl_id           28607 non-null    float64
28 contrib_factr_p2:id          4515 non-null    float64
29 units involved                147743 non-null   object
30 atd mode_category_metadata    147743 non-null   object
31 pedestrian_fl                 3505 non-null    object
32 motor vehicle fl              146634 non-null   object
33 motorcycle_fl                 3602 non-null    object
34 bicycle_fl                    2444 non-null    object
35 other fl                      4845 non-null    object
36 point                          145507 non-null   object
37 apd_confirmed_fatality        147750 non-null   object
38 apd_confirmed_death_count     147750 non-null   int64 int64
39 motor vehicle death count     147750 non-null   int64 int64
40 motor vehicle_serious_injury_count 147750 non-null   int64 int64
41 bicycle_death_count           147750 non-null   int64 int64
42 bicycle_serious_injury_count  147750 non-null   int64 int64
43 pedestrian_death_count        147750 non-null   int64
44 pedestrian_serious_injury_count 147750 non-null   object
45 motorcycle_death_count         147750 non-null   object
46 motorcycle_serious_injury_count 147750 non-null   int64 int64
47 other death count             147750 non-null   object
48 other_serious_injury_count    147750 non-null   object
49 onsys_fl                       147750 non-null   object
50 private_dr_fl                  147750 non-null   object
51 micromobility_serious_injury_count 147750 non-null   object
52 micromobility_death_count      147750 non-null   object
53 micromobility_fl               311 non-null     object
```

dtypes: float64(14), int64(17), object(23)

memory usage: 60.9+ MB

[147750 rowsx 54 columns]



Reproduction

Dataset statistics

Number of variables	54
Number of observations	147750
Missing cells	1725084
Missing cells (%)	21.6%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	60.9 MiB
Average record size in memory	432.0 B

Variable types

Numeric	17
Boolean	11
DateTime	2
Text	7
Unsupported	2
Categorical	15

Shruti Randive
NUID 002740632

For Example:

If we want the details of the street name column we can get it through the following data

street_name	
Text	
Distinct	4630
Distinct (%)	3.1%
Missing	2
Missing (%)	< 0.1%
Memory size	1.1 MiB



[More details](#)

Here we can see the details of the street name for the above data

Overview	Words	Characters	
Length		Characters and Unicode	Unique
Max length	41	Total characters	2208
Median length	40	Distinct characters	1370949
Mean length	9.2789682	Distinct categories	37
Min length	3	Distinct scripts	2
		Distinct blocks	1
The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.			
Unique	2208	?	Sample
Unique (%)	1.5%		1st row
			3707 MANCHACA RD
			2nd row
			PALM WAY TO MOPAC NB RAMP
			3rd row
			BALCONES CLUB DR
			4th row
			US0290
			5th row
			US0290

Data Staging using Talend

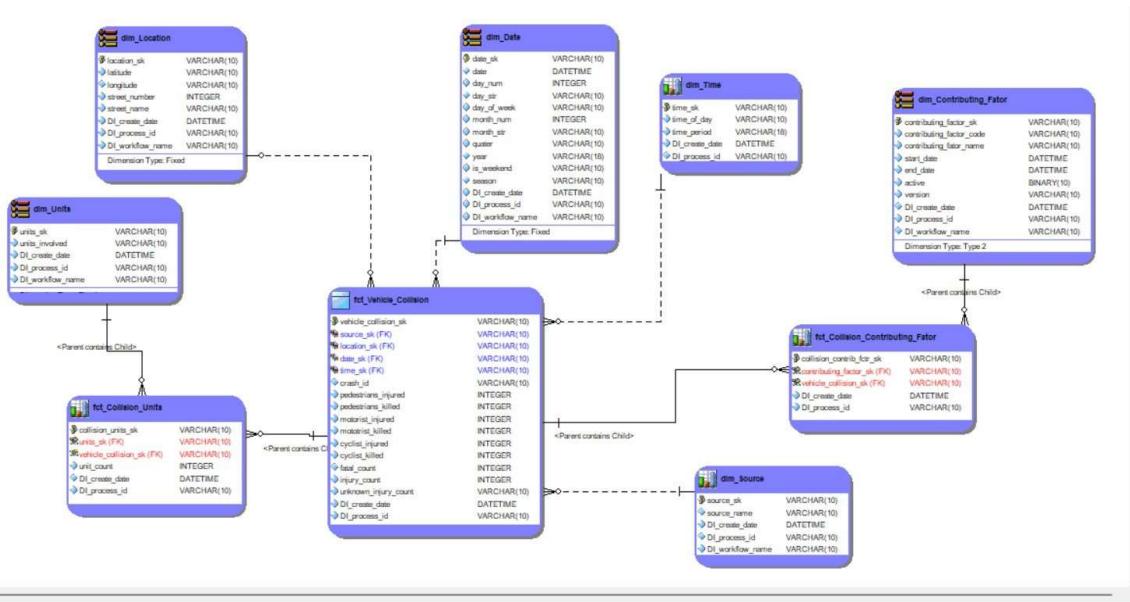
We are staging the source data using talend ETL pipelines for all the three datasets and storing the data into SQL server tables. We are going to use this staged data for transformations according to the business requirements.

Dimensional model:

We have outlined the process of creating a dimensional model consisting of both facts and dimensions, along with the necessary mapping documents and explanations of source-to-target column mappings.

Facts and Dimensions Creation:

- Identified the relevant facts and dimensions based on business requirements and data analysis.
- Created the necessary tables to represent these facts and dimensions.



We have created five dimensional tables and two fact tables along with a source table, all having a many to one relationships

Fact Table:

The fct_Vehicle_Collision table is one fact table in this model. It contains detailed information about individual vehicle collisions, such as:

- Location details (linked to the dim_Location table)
- Date and time details (linked to dim_Date and dim_Time tables)
- Units involved (linked to the dim_Units table)
- Contributing factors (linked to the fct_Collision_Contributing_Factor junction table and dim_Contributing_Factor table)
- Collusion details (e.g., crash_id, crash_id (FK), units involved, pedestrians injured/killed, motorists injured/killed, cyclists injured/killed, injury counts)
- Additional attributes like DI_create_date and DI_process_id

The fct_Collision_Contributing_Factor is the second fact table that we have created and has the following attributes:

- collision_contrib_fct_sk: Surrogate key for the fact table
- vehicle_collision_sk (FK): Foreign key referencing the Vehicle Collision fact table
- contributing_factor_sk (FK): Foreign key referencing the dim_Contributing_Factor dimension table
- DI_create_date: Date when the record was created
- DI_process_id: Process ID related to the record

The fct_Collision_Units is the third fact table that we have created and has the following attributes:

- collision_unit_sk: Surrogate key for the fact table
- unit_sk (FK): Foreign key referencing the Vehicle Collision fact table
- vehicle_collision_sk (FK): Foreign key referencing the Vehicle Collision fact table
- DI_create_date: Date when the record was created
- DI_process_id: Process ID related to the record
- unit_count : linked to the dim_Units table

Dimensional Tables:

dim_Location: This table stores location-related information, such as latitude, longitude, street number, and street name.

dim_Date: This table contains date-related attributes like date, day, month, year, quarter, and week.

dim_Time: This table holds time-related information, such as time, hour, and minute.

dim_Units: This table stores details about the units involved in the collision, such as the unit type and the process ID.

dim_Contributing_Factor: This table contains information about contributing factors related to the collision, such as the contributing factor code, factor name, start date, and end date.

dim_Source: This table has the source data which will be used for loading and processing data.

Mapping Document:

- Developed a comprehensive mapping document that clearly specifies the relationship between source and target columns.
- Documented each transformation applied during the ETL process, ensuring transparency and traceability.

The screenshot for the mapping document is as follows:

A	B	C	D	E	F	G
Final Column	Austin	Transformation	Chicago	Transformation	New York	Transformation
1 crash_id	Columns crash_id		Columns crash_record_id		Columns collision_id	
2 crash_date	crash_date		crash_date		crash_date	merge crash date and crash time
3 crash_time	crash_time	parse crash_date and get crash timein 24hr format	parse crash_date and get crash timein 24hr format	parse crash_time		
4 street number	street_nbr		street_no		doesn't have column	doesn't have column
5 street name	street_name		street_name		on_street_name	
6 latitude	latitude	will use street_nbr, street_nbr_2 column is null	latitude		location	split location and get latitude
7 longitude	longitude		longitude		location	split location and get longitude
8 contributing_factor_name	contributing_factor_name	with the help of contributing code we get the contributing factor	primary_contributory_cause +secondary_contributory_cause	concat these columns, normalize them	contributing_fact_vehicel1+contributing_fact_vehicel2+contributing_fact_vehicel3+contributing_fact_vehicel4+contributing_fact_vehicel5	concat these columns, normalize them with the help of document we can get code for each of the contributing factor name
9				with the help of document we can get code for each of the contributing factor name		
10						
11 contributing_factor_code	contrib_factr_p1_id+contrib_factr_p2_id	concat these columns and normalize			vehicle_type_code1+vehicle_type_code2+vehicle_type_code3+vehicle_type_code4+vehicle_type_code5	normalize them with the help of document we can get code for each of the contributing factor name
12 units_involved	units_involved		doesn't have column	doesn't have column	vehicle_type_code1+vehicle_type_code2+vehicle_type_code3+vehicle_type_code4+vehicle_type_code5	concat these columns, normalize them
13 unit_count		normalize the column 'units_involved' and then group by on crash_id which will gives us the unit count	no column specific to only units, num_units has information about number of pedestreians, vehicles and cyclists as well	no column specific to only units, num_units has information about number of pedestreians, vehicles and cyclists as well	vehicle_type_code1+vehicle_type_code2+vehicle_type_code3+vehicle_type_code4+vehicle_type_code5	concat these columns and groupby to count units
14 pedestrian_injured	pedestrian_serious_injury_count		doesn't have column	doesn't have column	number_of_pedestrian_injured	
15 pedestrian_killed	pedestrian_death_count		doesn't have column	doesn't have column	number_of_pedestrian_killed	
16 pedestrian_involved	pedestrian_death_count+pedestrian_serious_injury_count		doesn't have column	doesn't have column	number_of_pedestrian_injured+number_of_pedestrian_killed	

17	roadusers_involved	[motorcycle_death_count+motorcycle_serious_injury_count+motor_vehicle_death_count+motor_vehicle_serious_injury_count+bicycle_death_count+bicycle_serious_injury_count+motorcycle_serious_injury_count+motor_vehicle_death_count+bicycle_death_count]	doesn't have column	doesn't have column	number_of_cyclist_injured+number_of_cyclist_killed+number_of_motorist_injured+number_of_motorist_killed
18	motorist_injured		doesn't have column	doesn't have column	number_of_motorist_injured
19	motorist_killed		doesn't have column	doesn't have column	number_of_motorist_killed
20	cyclist_injured		doesn't have column	doesn't have column	number_of_cyclist_injured
21	cyclist_killed		doesn't have column	doesn't have column	number_of_cyclist_killed
22	fatal_count		injuries_fatal		no of person killed+ no of pedestrian killed+no of cyclist killed+no of motorist killed
23	Injury_count	[motorcycle_serious_injury_count+pedestrian_serious_injury_count+bicycle_serious_injury_count+motor_vehicle_serious_injury_count+other_serious_injury_count+micromobility_serious_injury_count]	Injuries non incapacitating+injuries_incapacitating+injuries reported not evident+injuries_no indication + injuries unknown	merge these columns	no of person injured+ no of pedestrian injured+no of cyclist injured+no of motorist injured
24	unknown_injury_count	no column	no column		no column

The Target columns are as follows:

- Crash_id
- Crash_date
- Crash_time
- Street number
- Street name
- Latitude
- Longitude
- Contributing_factor_name
- Contributing_factor_code
- Units_involved
- Unit_count
- Pedestrain_injured
- Pedestrian_killed
- Pedestrian_involved
- Roadusers_involved
- Motorist_injured
- Motorist_killed
- Cyclist_injured
- Cyclist_killed
- Fatal_count
- Injury_count
- Unknown_injury_count

Explanation:

crash_id: This column represents a unique identifier for each crash event. It is mapped directly from the source column "crash_id" in Austin, "crash_record_id" in Chicago, and "collision_id" in New York.

Shruti Randive
NUID 002740632

crash_date: This column stores the date of each crash event. The transformation involves parsing the "crash_date" from the source and merging it with the "crash_time" to create a unified datetime format in 24-hour format.

crash_time: Represents the time of the crash event. This is derived from parsing the "crash_date" in the source data to extract the time component.

street_number: Stores the street number where the crash occurred. The transformation involves selecting "street_nbr" in Austin and "street_no" in Chicago. For New York, there is no specific column provided.

street_name: Represents the name of the street where the crash occurred. This is directly mapped from the source column "street_name".

latitude: Stores the latitude coordinate of the crash location. This is obtained by splitting the "location" column in the source data to extract latitude information.

longitude: Represents the longitude coordinate of the crash location. Similar to latitude, this is obtained by splitting the "location" column in the source data.

contributing_factor_name: This column represents the primary contributing factor(s) to the crash event. The transformation involves concatenating and normalizing contributing factors from multiple source columns.

contributing_factor_code: Stores the code corresponding to each contributing factor. This is obtained by concatenating and normalizing contributing factor codes from multiple source columns.

units_involved: Represents the number of units (e.g., vehicles, pedestrians) involved in the crash event. This is obtained by concatenating and normalizing vehicle type codes from multiple source columns.

unit_count: Stores the count of units involved in each crash event. This is obtained by normalizing the "units_involved" column and performing a group-by operation on the crash ID.

pedestrian_injured: Stores the count of pedestrians injured in the crash event. pedestrian_killed:

Represents the count of pedestrians killed in the crash event.

Shruti Randive
NUID 002740632

pedestrian_involved: Represents the total count of pedestrians involved in the crash event (both injured and killed).

roadusers_involved: Represents the total count of road users involved in the crash event, including motorists, cyclists, and pedestrians.

motorist_injured: Stores the count of motorists injured in the crash event.

motorist_killed: Represents the count of motorists killed in the crash event.

cyclist_injured: Represents the count of cyclists injured in the crash event.

cyclist_killed: Stores the count of cyclists killed in the crash event.

fatal_count: Represents the total count of fatalities in the crash event, including pedestrians, cyclists, and motorists.

injury_count: Stores the total count of injuries in the crash event, including pedestrians, cyclists, and motorists.

unknown_injury_count: Represents the count of injuries with unknown severity or indication.

Transformation Explanations:

crash_id: No transformation is applied here; it directly maps from the source column in each city.

crash_date: The transformation involves parsing the "crash_date" from the source and merging it with the "crash_time" to create a unified datetime format in 24-hour format.

crash_time: This column is derived from parsing the "crash_date" in the source data to extract the time component.

street_number: In Austin, it uses the "street_nbr" column, and if "street_nbr_2" is null, it takes its value. In Chicago, it directly maps to the "street_no" column. For New York, there is no specific column provided.

street_name: No transformation is applied here; it directly maps from the source column.

latitude: The transformation involves splitting the "location" column in the source data to extract the latitude information.

longitude: Similarly, this column is derived from splitting the "location" column in the source data to extract the longitude information.

contributing_factor_name: The transformation involves concatenating and normalizing contributing factors from multiple source columns.

contributing_factor_code: This column is obtained by concatenating and normalizing contributing factor codes from multiple source columns.

units_involved: It represents the number of units involved in the crash event, obtained by concatenating and normalizing vehicle type codes from multiple source columns.

unit_count: The transformation involves normalizing the "units_involved" column and performing a group-by operation on the crash ID to count the units involved.

pedestrian_injured: No transformation is applied here, it directly maps from the source column.

pedestrian_killed: Similarly, this column directly maps from the source column.

pedestrian_involved: This column represents the total count of pedestrians involved in the crash event, obtained by adding pedestrian death and serious injury counts.

roadusers_involved: Represents the total count of road users involved in the crash event, including motorists, cyclists, and pedestrians. It's obtained by adding counts of motorcycle deaths, motorcycle serious injuries, motor vehicle deaths, motor vehicle serious injuries, bicycle deaths, and bicycle serious injuries.

motorist_injured: This column represents the count of motorists injured in the crash event, obtained by adding counts of motorcycle serious injuries and motor vehicle serious injuries.

motorist_killed: Similarly, this column represents the count of motorists killed in the crash event, obtained by adding counts of motorcycle deaths and motor vehicle deaths.

cyclist_injured: Represents the count of cyclists injured in the crash event, obtained from the count of bicycle serious injuries.

cyclist_killed: Similarly, this column represents the count of cyclists killed in the crash event, obtained from the count of bicycle deaths.

fatal_count: This column represents the total count of fatalities in the crash event, including pedestrians, cyclists, and motorists. It's obtained by summing up various death counts from different sources.

injury_count: Represents the total count of injuries in the crash event, including pedestrians, cyclists, and motorists. It's obtained by summing up various serious injury counts from different sources.

unknown_injury_count: Represents the count of injuries with unknown severity or indication, obtained by merging multiple columns indicating different types of unknown injuries.

NULL values:

Regarding the Null values, we will be handling it during visualization as we will be getting a clear picture of the null count.

Part 2 :

1) Loading data into stage table and their row counts

Stg_Austin:



Shruti Randive
NUID 002740632

```

SELECT * FROM [Vehicle_Collision].[dbo].[stg_austin_collision]
SELECT * FROM [Vehicle_Collision].[dbo].[stg_chicago_collision]
SELECT * FROM [Vehicle_Collision].[dbo].[stg_newyork_collision]

SELECT * FROM [Vehicle_Collision].[dbo].[austin_transformation]
SELECT * FROM [Vehicle_Collision].[dbo].[chicago_transformation]
SELECT * FROM [Vehicle_Collision].[dbo].[newyork_transformation]

SELECT * FROM [Vehicle_Collision].[dbo].[austin_contribution_factor]
SELECT * FROM [Vehicle_Collision].[dbo].[chicago_contribution_factor]
SELECT * FROM [Vehicle_Collision].[dbo].[newyork_contribution_factor]

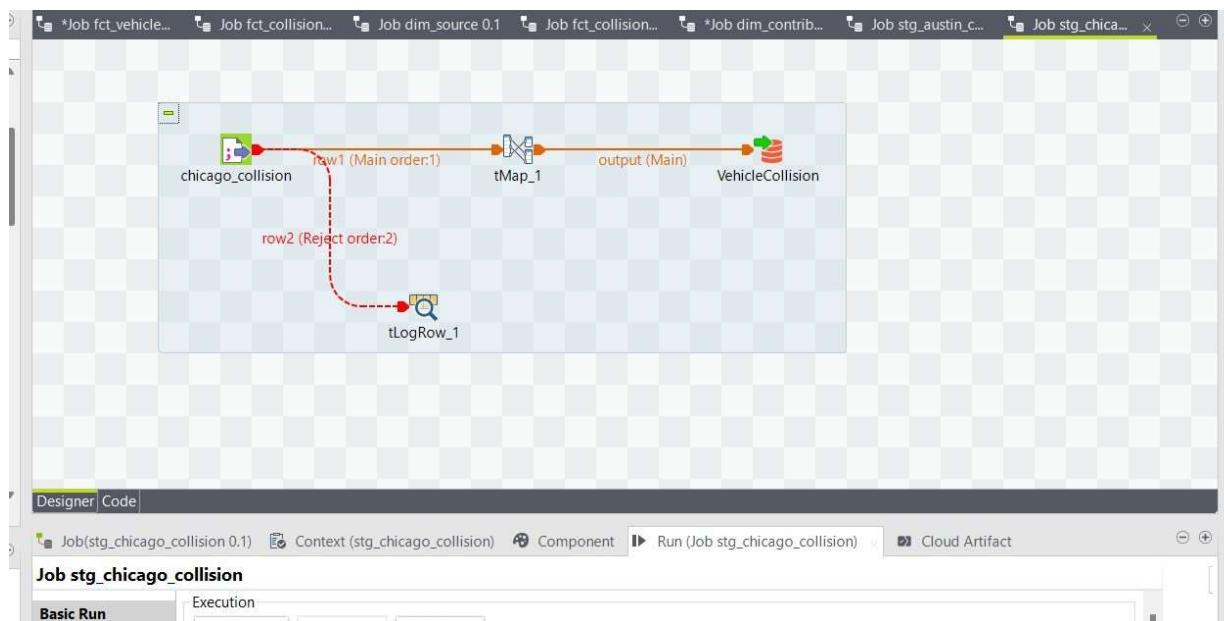
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Location]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Units]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Date]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Time]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Source]

CREATE TABLE [Vehicle_Collision].[dbo].[dim_Contributing_Factor](
    [contributing_factor_sk] [int] IDENTITY(1,1) NOT NULL,
    [contributing_factor_code] [varchar](150) NULL,
    [contributing_factor_name] [varchar](61) NULL,
    [scd_start] [datetime] NOT NULL,
    [scd_end] [datetime] NULL
)

```

crash_id	crash_date	crash_time	case_id	rpt_latitude	rpt_longitude	rpt_block_num	rpt_street_pfx	rpt_street_name	rpt_street_sfx	cash_speed_limit	road_constr_zone_S	latitude
14924014	N	02/13/2018 02:45:00 PM	14:45:00	160441073		6700		NOT REPORTED	55	N	30.23404582	
14933410	N	03/08/2018 04:23:00 PM	16:23:00	160681270		13800		NOT REPORTED	70	N	30.42486597	
3	14967979	N	03/11/2018 08:48:00 AM	08:48:00	160710490	10500		STAKE PLAINS	DR	30	30.4870376	
4	14993533	N	03/25/2018 07:20:00 AM	07:20:00	160850345	200	E	ANDERSON	LN	65	30.34323554	
5	15036491	N	04/14/2018 11:47:00 PM	23:47:00	161051957	3500	S	US 183 SB	HWY	55	30.19826831	
6	1497966	N	03/05/2018 10:40:00 PM	22:40:00	160652058	7600	E	RIVERBODE	DR	45	30.219193	
7	1500700	N	04/03/2018 12:25:00 AM	00:25:00	16094079	1400	E	OLTORF	ST	35	30.23386223	
8	1503674	N	04/17/2018 02:24:00 AM	14:24:00	161080932	13320	N	N FM 620	RD	-1	30.47039914	
9	15049470	N	04/19/2018 10:29:00 PM	22:29:00	161107182	3300		DAVE LN	LN	-1	30.19349307	
10	15049902	N	04/20/2018 09:55:00 AM	09:55:00	161110537	5300		NOT REPORTED	60	N	30.31068138	
...	14967979	N	04/26/2018 07:41:00 AM	07:41:00	160871277	7300	E	SHREWDINE	RD	1	30.49164279	

Stg_Chicago:



Shruti Randive
NUID 002740632

The screenshot shows the Microsoft SQL Server Management Studio interface. The Object Explorer on the left lists various database objects including tables, views, and stored procedures. The central pane displays a T-SQL script for creating a dimension table:

```

SELECT * FROM [Vehicle_Collision].[dbo].[stg_austin_collision]
SELECT * FROM [Vehicle_Collision].[dbo].[stg_chicago_collision]
SELECT * FROM [Vehicle_Collision].[dbo].[stg_newyork_collision]

SELECT * FROM [Vehicle_Collision].[dbo].[austin_transformation]
SELECT * FROM [Vehicle_Collision].[dbo].[chicago_transformation]
SELECT * FROM [Vehicle_Collision].[dbo].[newyork_transformation]

SELECT * FROM [Vehicle_Collision].[dbo].[austin_contribution_factor]
SELECT * FROM [Vehicle_Collision].[dbo].[chicago_contribution_factor]
SELECT * FROM [Vehicle_Collision].[dbo].[newyork_contribution_factor]

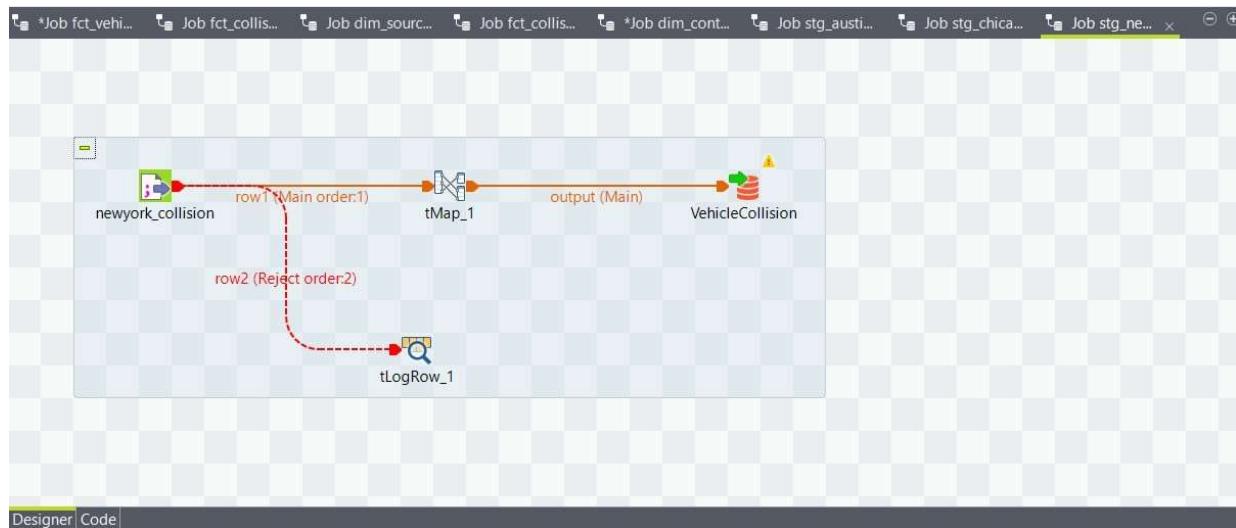
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Location]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Units]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Date]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Time]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Source]

CREATE TABLE [Vehicle_Collision].[dbo].[dim_Contributing_Factor](
    [contributing_factor_sk] [int] IDENTITY(1,1) NOT NULL,
    [contributing_factor_code] [varchar](150) NULL,
    [contributing_factor_name] [varchar](61) NULL,
    [scd_start] [datetime] NOT NULL,
    [scd_end] [datetime] NULL
)

```

The results pane below shows the output of the query, which includes 817,723 rows of data from the Vehicle_Collision table.

Stg_NewYork:



```

SELECT * FROM [Vehicle Collision].[dbo].[stg_austin_collision]
SELECT * FROM [Vehicle Collision].[dbo].[stg_chicago_collision]
SELECT * FROM [Vehicle Collision].[dbo].[stg_newyork_collision]

SELECT * FROM [Vehicle Collision].[dbo].[austin_transformation]
SELECT * FROM [Vehicle Collision].[dbo].[chicago_transformation]
SELECT * FROM [Vehicle Collision].[dbo].[newyork_transformation]

SELECT * FROM [Vehicle Collision].[dbo].[austin_contribution_factor]
SELECT * FROM [Vehicle Collision].[dbo].[chicago_contribution_factor]
SELECT * FROM [Vehicle Collision].[dbo].[newyork_contribution_factor]

SELECT * FROM [Vehicle Collision].[dbo].[dim_Location]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Units]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Date]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Time]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Source]

CREATE TABLE [Vehicle Collision].[dbo].[dim_Contributing_Factor](
[contributing_factor_sk] [int] IDENTITY(1,1) NOT NULL,
[contributing_factor_code] [varchar](150) NULL,
[contributing_factor_name] [varchar](61) NULL,
[scd_start] [datetime] NOT NULL,
[scd_end] [datetime] NULL
)

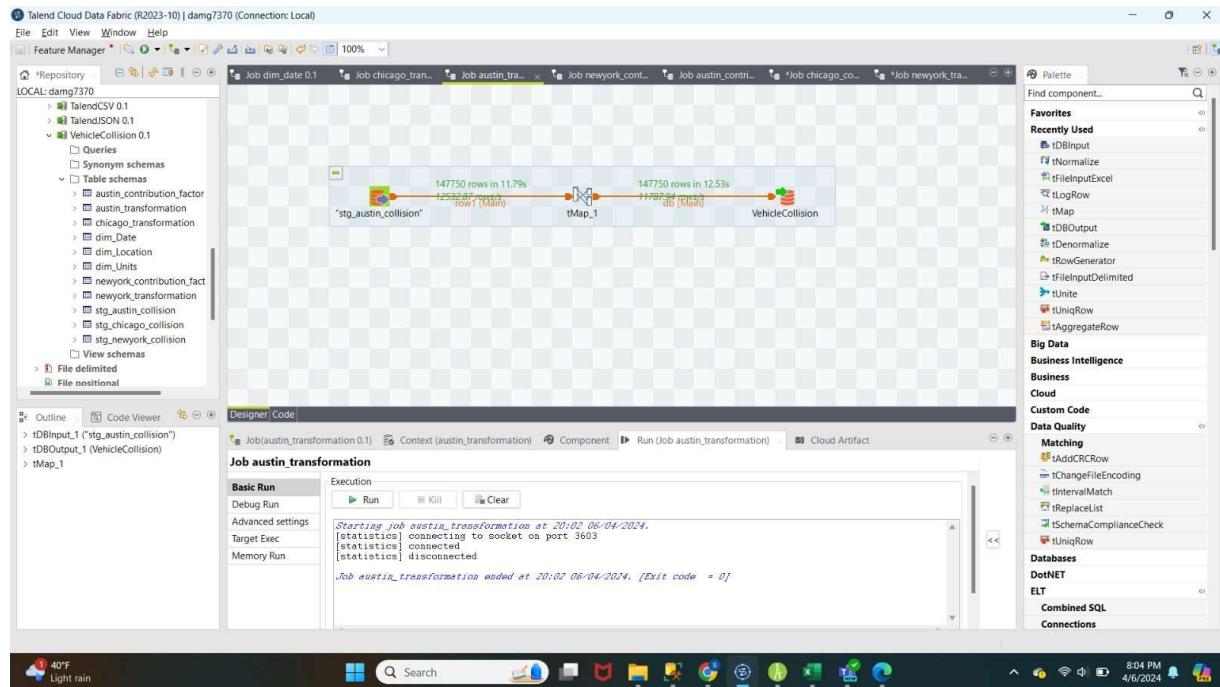
```

	CRASH_DATE	CRASH_TIME	BOROUGH	ZIP_CODE	LATITUDE	LONGITUDE	LOCATION	ON_STREET_NAME	CROSS_STREET_NAME	OFF_STREET_NAME	NUMBER_OF_PERSONS_INJURED
1	6/22/2021	11:00	BROOKLYN	11205	40.696053	-73.978939	(40.696053,-73.978935)	BURKE AVENUE	RADCLIFF AVENUE	MONUMENT WALK	0
2	6/22/2021	15:25	BRONX	10469	40.871315	-73.880379	(40.871315,-73.880374)				1
3	6/22/2021	17:00	BROOKLYN	11220	40.637759	-74.000721	(40.637759,-74.002712)	56 STREET	8 AVENUE		0
4	6/10/2021	20:00	MANHATTAN	10026	40.800392	-73.995461	(40.800392,-73.95461)				1842 7 AVENUE
5	6/22/2021	15:00			4.062165	-7415.741	(40.62165,-74.15741)				2040 FOREST AVENUE
6	6/22/2021	18:00	MANHATTAN	10013	40.72554	-7400.757	(40.72554,-74.00757)	SPRING STREET	HUDSON STREET		0
7	6/22/2021	7:30			4.057741	-7396.237	(40.57741,-73.96237)	BRIGHTON BEACH AVENUE			0
8	6/10/2021	17:30	QUEENS	11412	40.099406	-73.374953	(40.099406,-73.374953)	202 STREET	115 AVENUE		0
9	6/22/2021	0:01	BRONX	10460	40.804377	-73.889956	(40.804377,-73.889956)		PROSPECT AVENUE	ELSMERE PLACE	0
10	6/22/2021	12:59	QUEENS	11432	40.701612	-73.799835	(40.701612,-73.799835)				89-20 161 STREET
					40.686177	-73.995514	(40.686177,-73.995514)	17 ST			n

Query executed successfully.

2) Loading the transformed data

We did transformation workflows for all the three dataset i.e. Austin, Chicago and New York. Below are the screenshots of the workflows.



Shruti Randive

NUID 002740632

S TalendCloud DataF bric(R2023-10)Finalproject(Connection:Rutuja)

File Edit Vie Window Help

Featu-eManager tli U Job Repository JobDi JobTr JobTransf JobTran JobDm JobNewJob JobFac JobBing

LOCALFinal vtaJob A..Signs v ulistandard Austin_Contributing_Factor0.1 Chnco_Contributing_Factor0.1 Dim_Dat 0.1 Dim_Location0.1 ;Dim_Lime 0.1 ;Dim_Utito.1 ;Fact_Vehicle_Collision0.1 ;Motor_Vehicle 0.1 ;B_NewYork_Contributing_Factor0.1 ;stg_Au_stm0.1 ;stg_Chicago0.1 ;stg_NewYor.0.1 -transform_Im Austin0.1 -random_Im Chicago 0.1 random_Ima_N_wYork 0.1 rInfrom_line0.1 > JobletDesign1 v Co'itext1 Austin0.1 conn:uiling_factor_r1odifed1

Cmh_Time0.1 d rriat,0N

Job Transform Time Chicago J.1.1 Context[Transform Time* Chicago] +9 Component Run [Job Transform Time Chicago]

Job Transform_Time_Chicago

E cul:im

Run Clear

DebugRun AdvancedEUsing Targetfile: MemoryRun

St*Trnig job Transform_Ti.w_Cfl- "9.0 t 16.5.07.01.702.J Targetfile: [statistics] connecting to socket on port 3382 [outt::atc::a] connected [statistics] disconnected

Job Transform_Time_Chicago ended at 16:53 07/04/2024. [Exit code = 0]

OutlinE CodeViewer

LineLimit wrpcp

B(z) Paette S@ i'a Favorites Re:centlyUsed tlogRow

'\$tDBOutput ntDBInput PtUniqRow ..._Unite ftnormalize 'tFileInputExcel lJavaRow tFileInputDelimit... tRowGenerator \$tDBSCD Big D*t Business Intelligence Business Cloud

jiltGroovy ...GroovyFile; :luava CtlJavaFile; lJavaRow tlibraryload :SetDynamicS... ::seGlobalVlr DAtaQuality Matching UIAddCRoN ;tehangelleEnco... _11InternalMatch BtReplace.ist

6:02 PM 4/7/2024

S TalendCloud DataF bric(R2023-10)Finalproject(Connection:Rutuja)

File Edit Vie Window Help

Featu-eManager tli U Job Repository JobDi JobTr JobTransf JobTran JobDm JobNewJob JobFac JobBing

LOCALFinal vtaJob A..Signs v ulistandard Austin_Contributing_Factor0.1 Chnco_Contributing_Factor0.1 Dim_Dat 0.1 Dim_Location0.1 ;Dim_Lime 0.1 ;Dim_Utito.1 ;Fact_Vehicle_Collision0.1 ;Motor_Vehicle 0.1 ;B_NewYork_Contributing_Factor0.1 ;stg_Au_stm0.1 ;stg_Chicago0.1 ;stg_NewYor.0.1 -transform_Im Austin0.1 -transform_Im Chicago 0.1 random_Ima_N_wYork 0.1 rInfrom_line0.1 > JobletDesign1 v Co'itext1 Austin0.1 conn:uiling_factor_r1odifed1

Cmh_Time0.1 d rriat,0N

Job Transform Time NewYork0.1 Context[Transform Time NewYork] +9 Component Run [Job Transform Time NewYork]

Job Transform_Time_NewYork

E cul:im

Run Clear

DebugRun AdvancedEUsing Targetfile: MemoryRun

707432 103/05/2024 02:50 PK -CHURCH AV:NITE 40 72225 -73 9 40 707432 03/05/2024 02:50 PK -CCEAN ST:ST 40,58657 73,9 40 707432 03/05/2024 02:50 PK SUTPHIN BOULEVARD 40 E,80477 -73,7 40 707432 03/05/2024 05:00 PK SUNNY MAH:S MILJU:: 40 E,10781, -73,9 40 707432 03/05/2024 04:10 PK 40 E,10781, -73,7 40 707432 03/05/2024 02:30 PK MILLER AV:ND'E 40 E,779, -73,8 40 707432 03/05/2024 08:00 AM EDSILL AV:ND'E 40,706,512 -73,8

[statistics] disconnected

Job Transform Time NewYork ended at 12:45 07/04/2024 [Exit code = 0]

OutlinE CodeViewer

LineLimit wrpcp

B(z) Paette S@ i'a Favorites Re:centlyUsed tlogRow

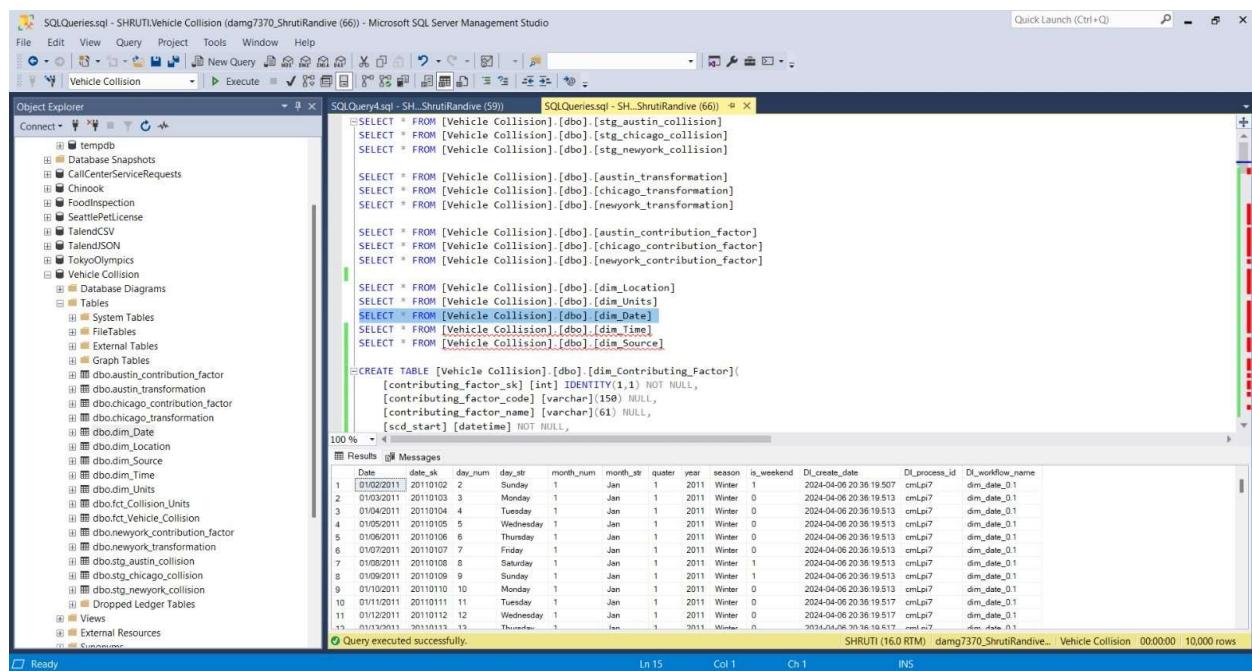
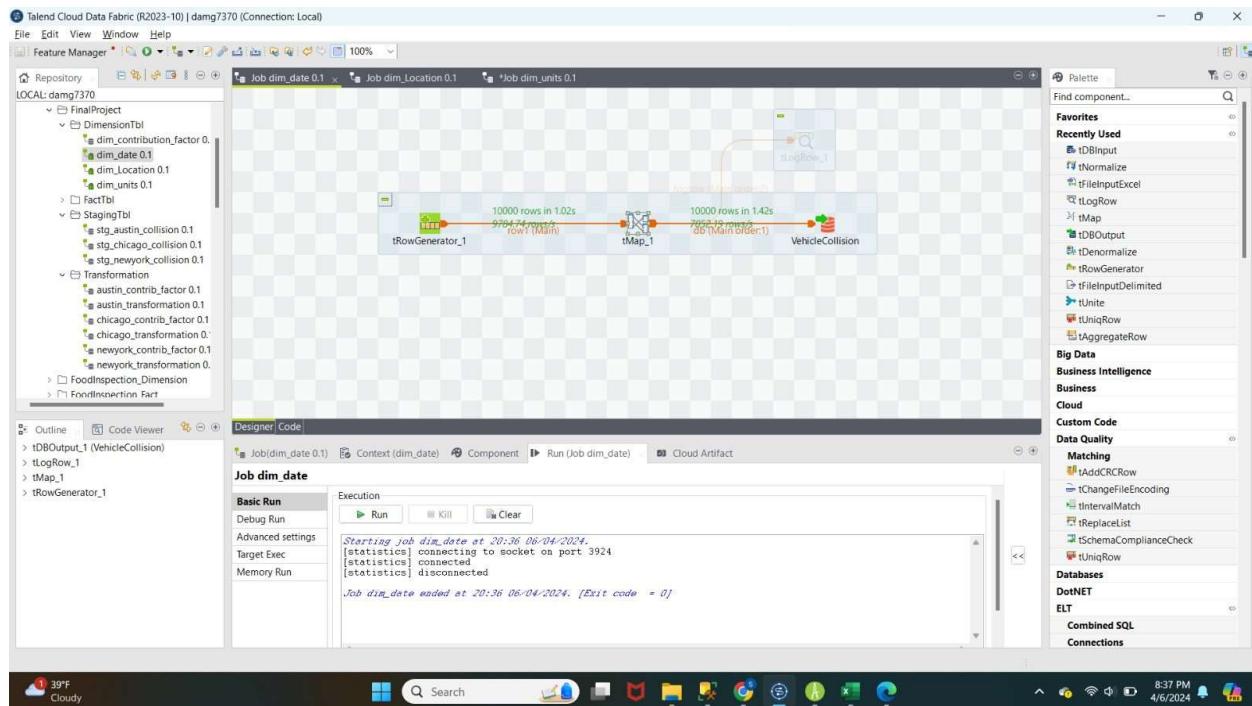
'\$tDBOutput ntDBInput PtUniqRow ..._Unite ftnormalize 'tFileInputExcel lJavaRow tFileInputDelimit... tRowGenerator \$tDBSCD Big D*t Business Intelligence Business Cloud

jiltGroovy ...GroovyFile; :luava CtlJavaFile; lJavaRow tlibraryload :SetDynamicS... ::seGlobalVlr DAtaQuality Matching UIAddCRoN ;tehangelleEnco... _11InternalMatch BtReplace.ist

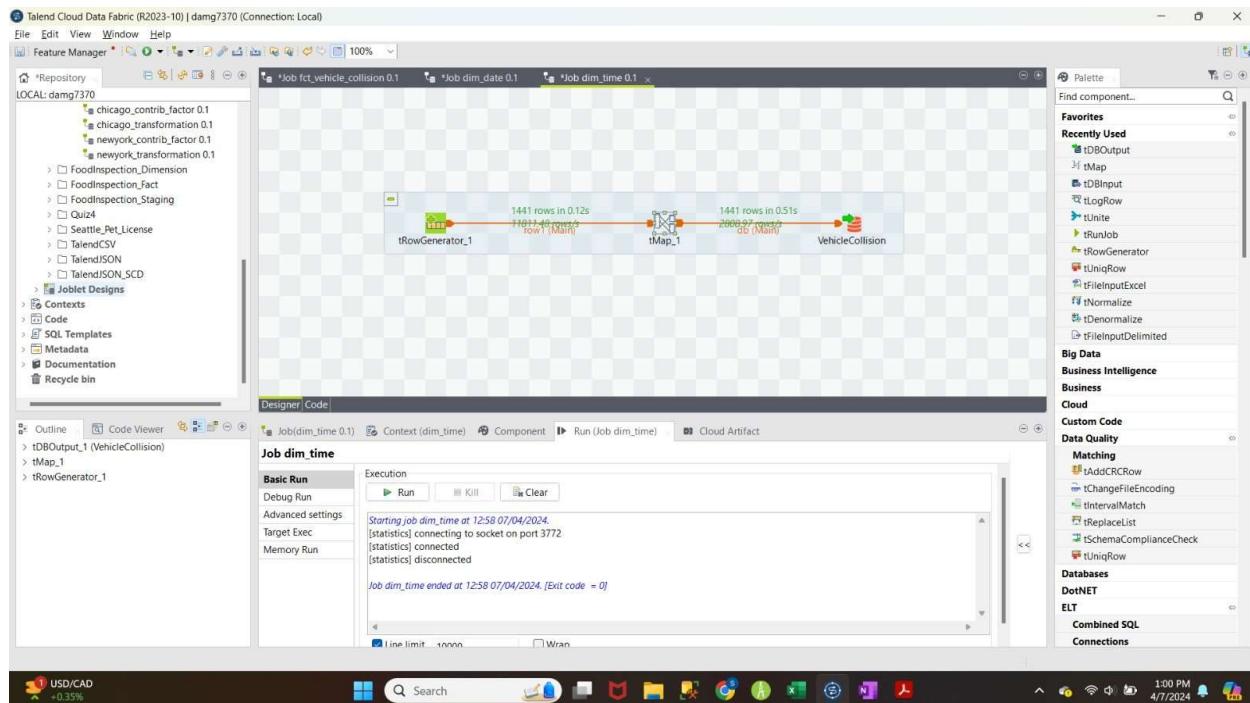
6:02 PM 4/7/2024

4)ETL for loading Dimensional tables:

Dim_Date:



Dim_Time:



SQLQueries.sql - SHRUTI.Vehicle Collision (damg7370_ShrutiRandive (66)) - Microsoft SQL Server Management Studio

File Edit View Query Project Tools Window Help

Object Explorer

SQLQuery4.sql - SH..ShrutiRandive (59) SQLQuery5.sql - SH..ShrutiRandive (66) ▾

```

SELECT * FROM [Vehicle_Collision].[dbo].[stg_austin_collision]
SELECT * FROM [Vehicle_Collision].[dbo].[stg_chicago_collision]
SELECT * FROM [Vehicle_Collision].[dbo].[stg_newyork_collision]

SELECT * FROM [Vehicle_Collision].[dbo].[austin_transformation]
SELECT * FROM [Vehicle_Collision].[dbo].[chicago_transformation]
SELECT * FROM [Vehicle_Collision].[dbo].[newyork_transformation]

SELECT * FROM [Vehicle_Collision].[dbo].[austin_contribution_factor]
SELECT * FROM [Vehicle_Collision].[dbo].[chicago_contribution_factor]
SELECT * FROM [Vehicle_Collision].[dbo].[newyork_contribution_factor]

SELECT * FROM [Vehicle_Collision].[dbo].[dim_Location]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Units]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Date]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Time]
SELECT * FROM [Vehicle_Collision].[dbo].[dim_Source]

```

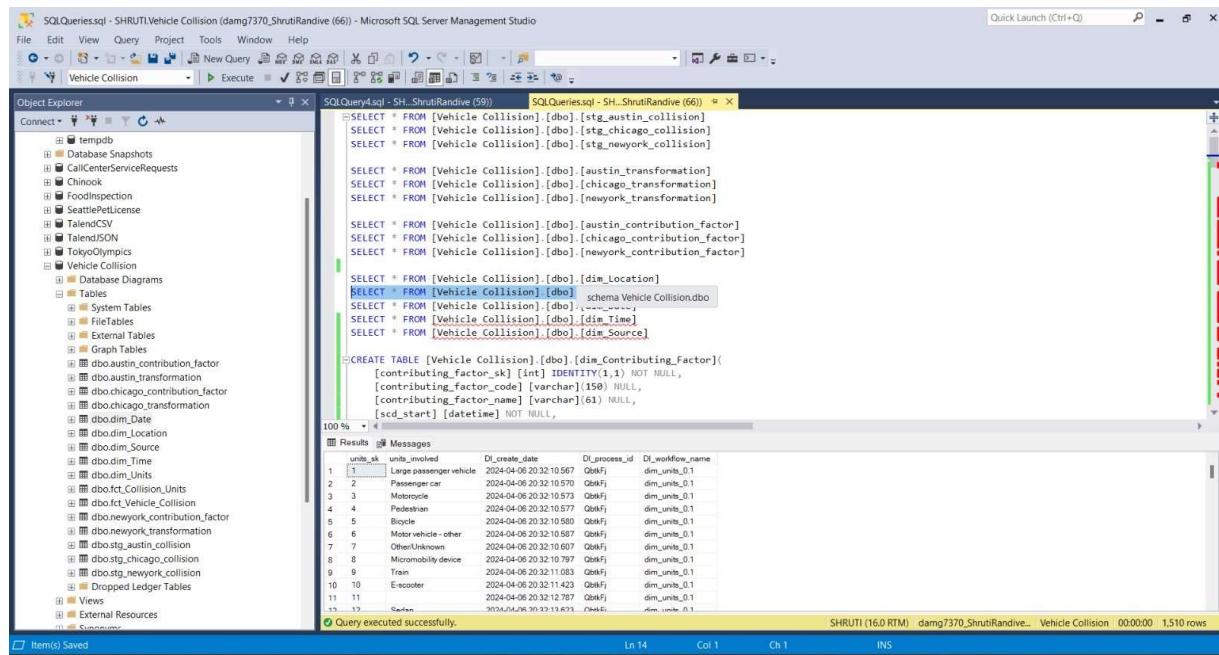
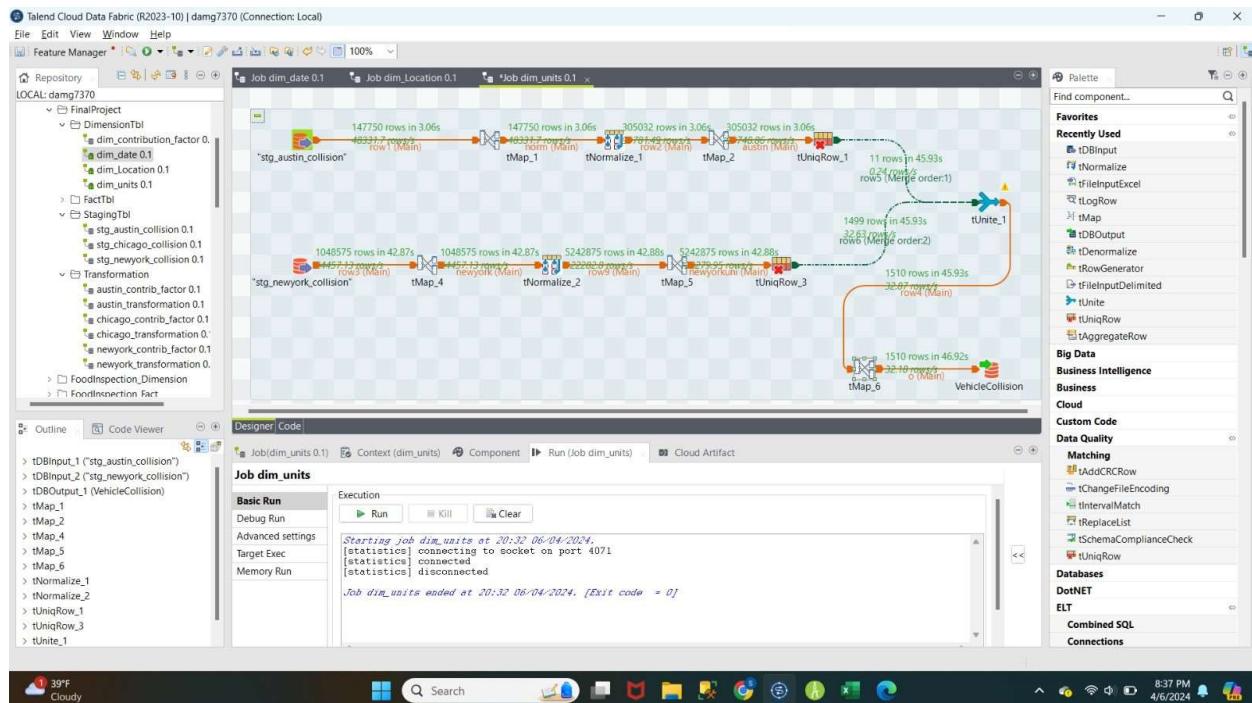
CREATE TABLE [Vehicle_Collision].[dbo].[dim_Contributing_Factor](
 [contributing_factor_sk] [int] IDENTITY(1,1) NOT NULL,
 [contributing_factor_code] [varchar](150) NULL,
 [contributing_factor_name] [varchar](61) NULL,
 [scd_start] [datetime] NOT NULL,

time_sk	time_of_day	time_of_period	di_create_date	di_process_id	di_workflow_name
1	0	12:00:00	Noon	2024-07-12 58:59:450	BFRnp dim_time_0.1
2	1	12:01:00	Noon	2024-07-12 58:58:457	BFRnp dim_time_0.1
3	2	12:02:00	Noon	2024-07-12 59:58:457	BFRnp dim_time_0.1
4	3	12:03:00	Noon	2024-07-12 59:59:457	BFRnp dim_time_0.1
5	4	12:04:00	Noon	2024-07-12 59:58:457	BFRnp dim_time_0.1
6	5	12:05:00	Noon	2024-07-12 59:58:457	BFRnp dim_time_0.1
7	6	12:06:00	Noon	2024-07-12 59:59:457	BFRnp dim_time_0.1
8	7	12:07:00	Noon	2024-07-12 58:58:457	BFRnp dim_time_0.1
9	8	12:08:00	Noon	2024-07-12 58:58:457	BFRnp dim_time_0.1
10	9	12:09:00	Noon	2024-07-12 58:58:457	BFRnp dim_time_0.1
11	10	12:10:00	Noon	2024-07-12 58:58:457	BFRnp dim_time_0.1
12	11	12:11:00	Noon	2024-07-12 58:58:457	BFRnp dim_time_0.1

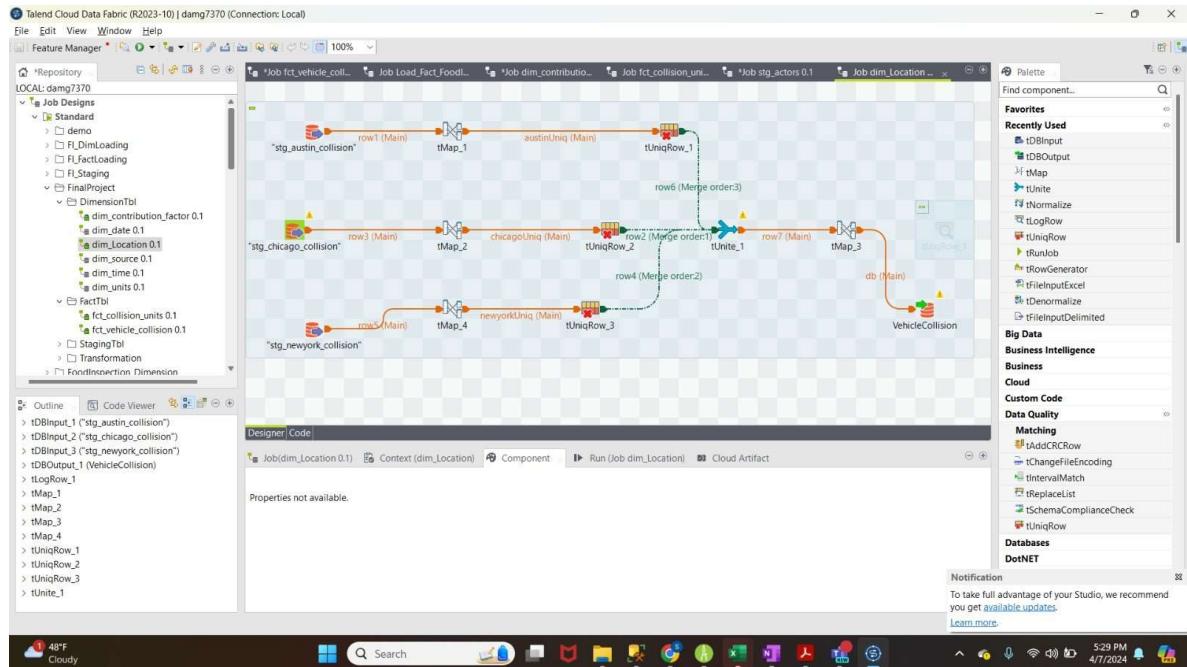
Query executed successfully.

SHRUTI (16.0 RTM) damg7370_ShrutiRandive... Vehicle Collision 00:00:00 1,441 rows

Dim_Units:



Dim_Location:



The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. The Object Explorer on the left shows the database structure, including tables like dim_Location, dim_Contributing_Factor, and dim_Collision_Units. The central pane displays a query results grid. The query itself is as follows:

```

SELECT * FROM [Vehicle Collision].[dbo].[stg_austin_collision]
SELECT * FROM [Vehicle Collision].[dbo].[stg_chicago_collision]
SELECT * FROM [Vehicle Collision].[dbo].[stg_newyork_collision]

SELECT * FROM [Vehicle Collision].[dbo].[austin_transformation]
SELECT * FROM [Vehicle Collision].[dbo].[chicago_transformation]
SELECT * FROM [Vehicle Collision].[dbo].[newyork_transformation]

SELECT * FROM [Vehicle Collision].[dbo].[austin_contribution_factor]
SELECT * FROM [Vehicle Collision].[dbo].[chicago_contribution_factor]
SELECT * FROM [Vehicle Collision].[dbo].[newyork_contribution_factor]

SELECT * FROM [Vehicle Collision].[dbo].[dim_Location]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Units]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Date]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Time]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Source]

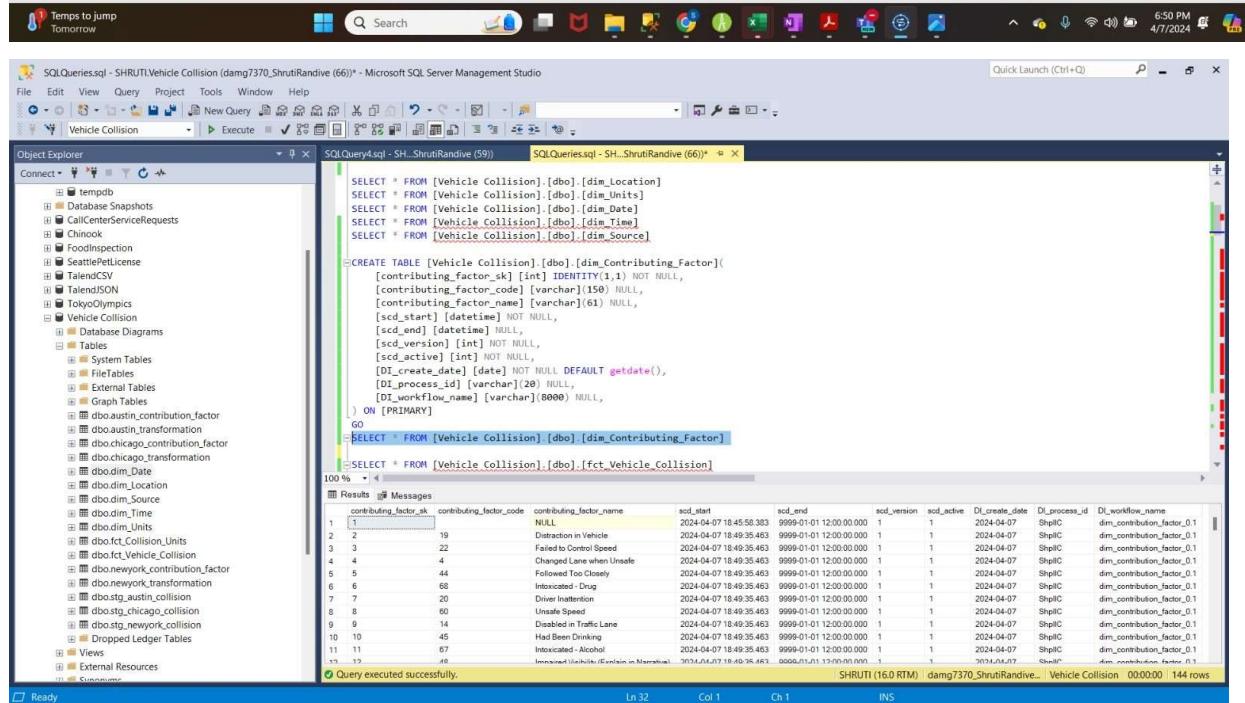
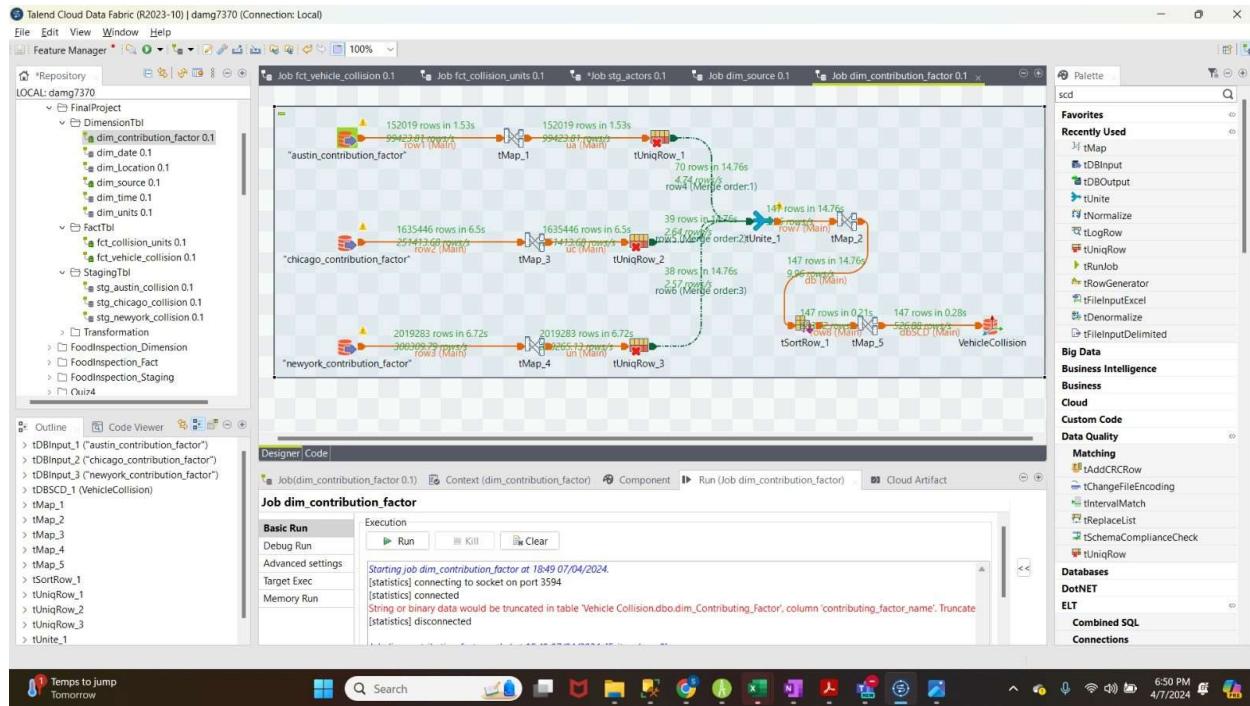
CREATE TABLE [Vehicle Collision].[dbo].[dim_Contributing_Factor](
    [contributing_factor_sk] [int] IDENTITY(1,1) NOT NULL,
    [contributing_factor_code] [varchar](150) NULL,
    [contributing_factor_name] [varchar](61) NULL,
    [scd_start] [datetime] NOT NULL,
    CONSTRAINT [PK_dim_Contributing_Factor] PRIMARY KEY CLUSTERED ([contributing_factor_sk])
)

```

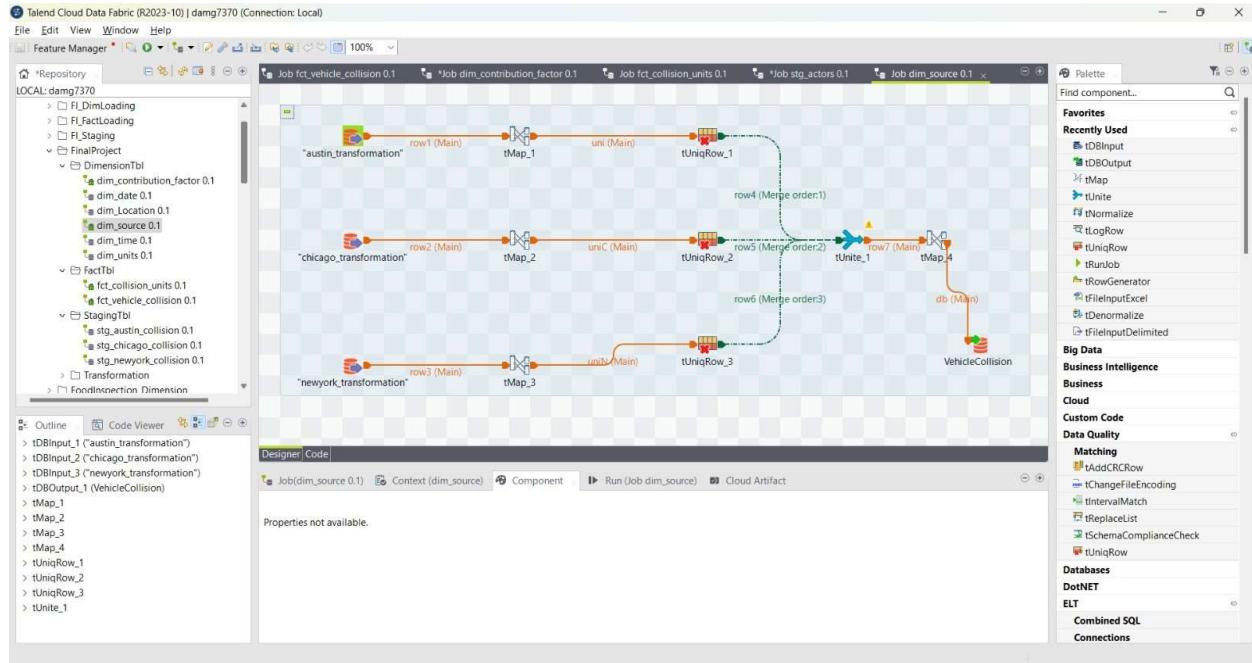
The results grid shows 598,899 rows of data, including columns like location_sk, address, street_number, street_name, dim_create_date, dim_process_id, and dim_workflow_name. The data includes various addresses like DAMEN AVE, CLARK ST, CERMACK RD, 94TH ST, 127TH ST, NARPMANSETT AVE, 47TH ST, LAKE SHORE DR NW, WELLINGTON AVE, FLETCHER ST, DEVON AVE, and BARRY AVE, spanning from April 2024 to May 2024.

Dim_Contributing_Factor:

We did Contributing_Factor workflows for all the three dataset i.e. Austin, Chicago and New York.
This is our SCD type 2, columns are added as required.



Dim_Source:



The screenshot shows Microsoft SQL Server Management Studio (SSMS) with a query window titled "SQLQueries.sql - SHRUTI\Vehicle Collision (damg7370_Shru... (66))". The Object Explorer on the left shows the database structure, including tables like tempdb, CallCenterServiceRequests, Chinook, FoodInspection, SeattlePetLicense, TalendCSV, TalendJSON, TokyoOlympics, Vehicle Collision, and various dimension and fact tables. The query window contains T-SQL code for creating a dimension table "dim_Contributing_Factor" and inserting data from three source tables: "stg_austin_collision", "stg_chicago_collision", and "stg_newyork_collision". The results pane shows the inserted data.

```

SELECT * FROM [Vehicle Collision].[dbo].[stg_austin_collision]
SELECT * FROM [Vehicle Collision].[dbo].[stg_chicago_collision]
SELECT * FROM [Vehicle Collision].[dbo].[stg_newyork_collision]

SELECT * FROM [Vehicle Collision].[dbo].[austin_transformation]
SELECT * FROM [Vehicle Collision].[dbo].[chicago_transformation]
SELECT * FROM [Vehicle Collision].[dbo].[newyork_transformation]

SELECT * FROM [Vehicle Collision].[dbo].[dim_Location]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Units]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Date]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Time]
SELECT * FROM [Vehicle Collision].[dbo].[dim_Source]

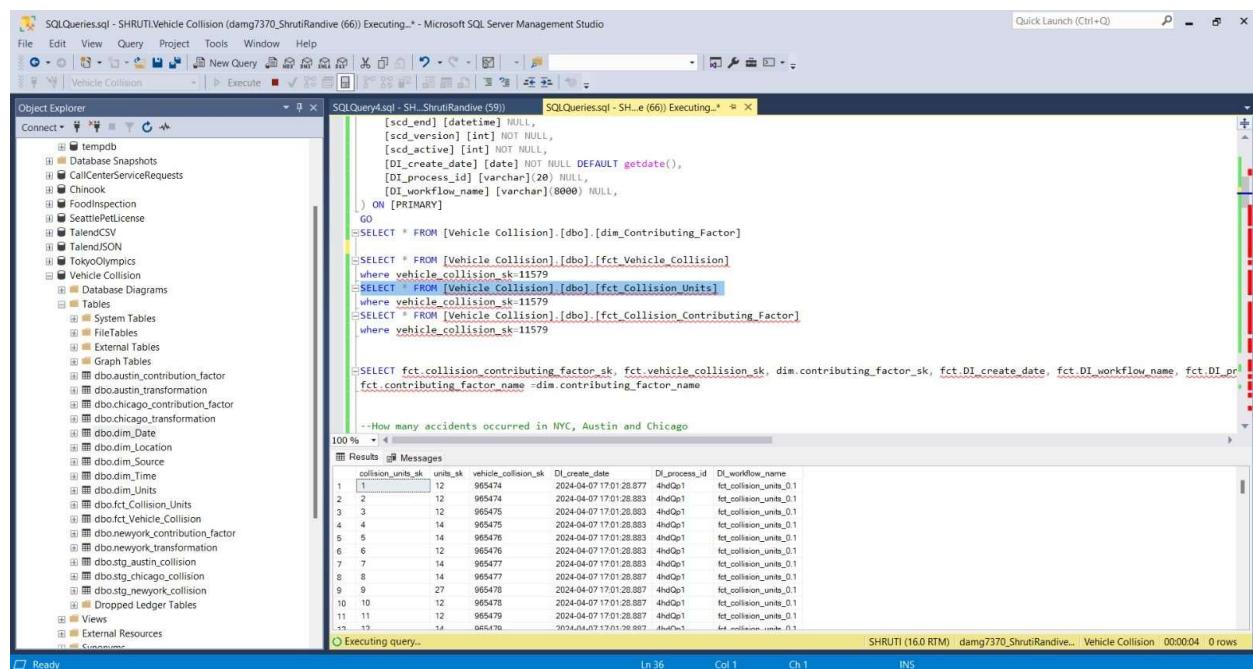
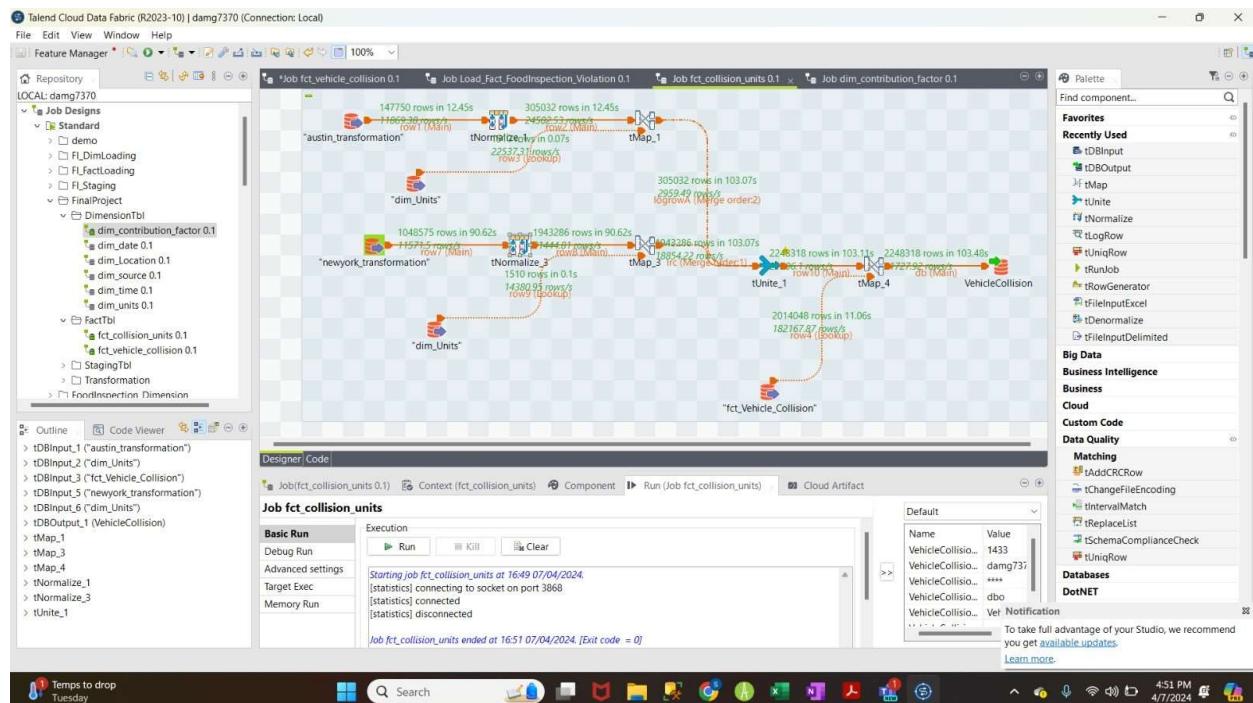
CREATE TABLE [Vehicle Collision].[dbo].[dim_Contributing_Factor](
    [contributing_factor_sk] [int] IDENTITY(1,1) NOT NULL,
    [contributing_factor_code] [varchar](150) NULL,
    [contributing_factor_name] [varchar](61) NULL,
    [scd_start] [datetime] NOT NULL,
    [scd_end] [datetime] NULL
)

source_sk source_name DI_create_date DI_process_id DI_workflow_name
1 austin 2024-04-07 10:53:30.627 2YNvzr dim_source_0_1
2 chicago 2024-04-07 10:53:31.880 2YNvzr dim_source_0_1
3 newyork 2024-04-07 10:53:38.307 2YNvzr dim_source_0_1

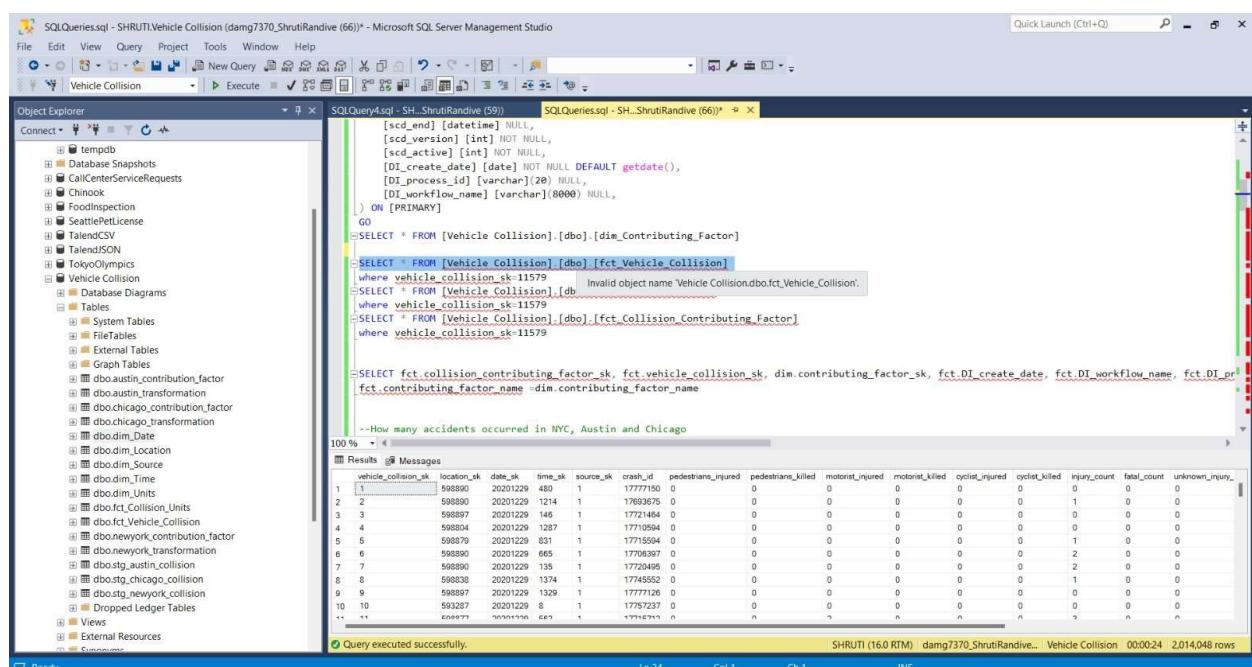
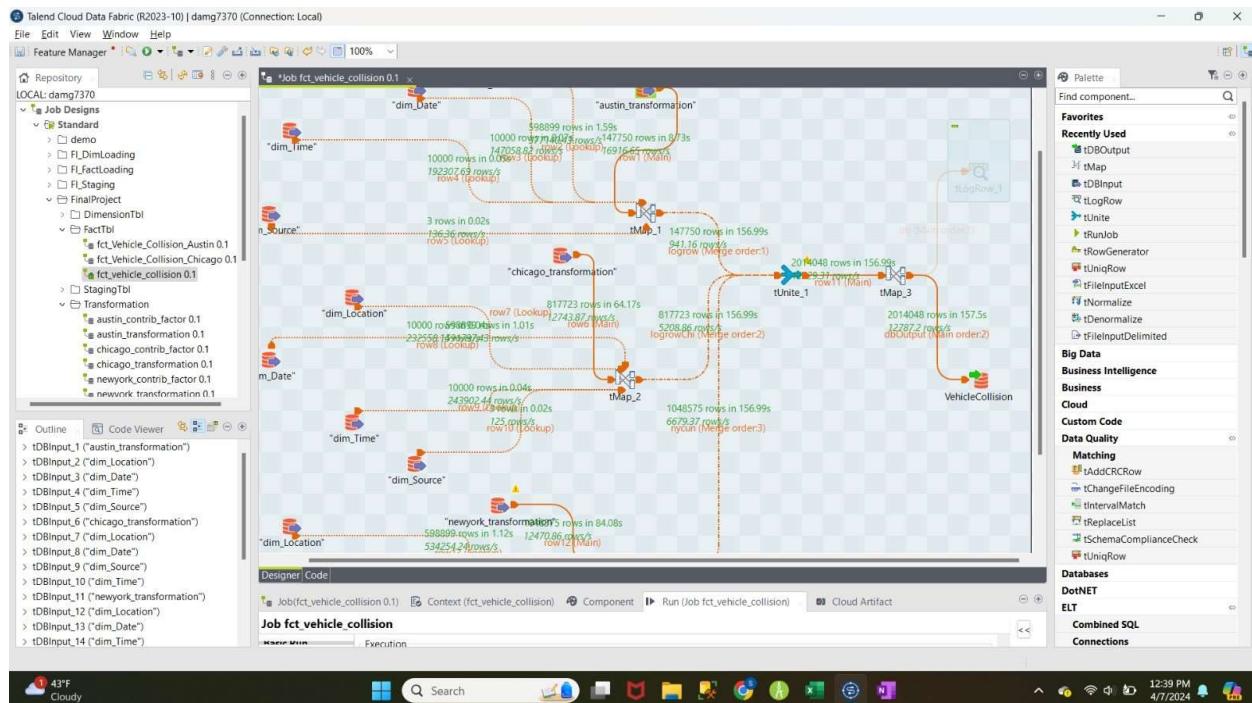
```

4)Fact Tables

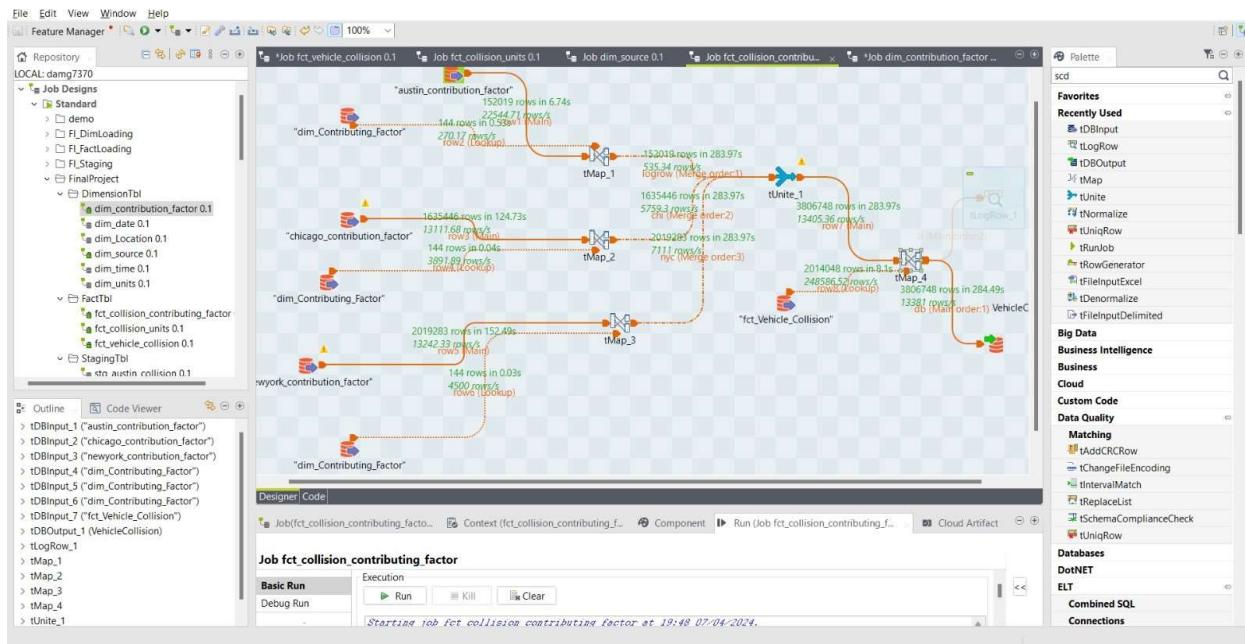
Fact_Collision_Units



Fact_Vehicle_Collision



Fact_Collision_Contribution_Factor



SQLQueries.sql - SHRUTIVehicleCollision (damg7370_ShruviRandive (66)) - Microsoft SQL Server Management Studio

File Edit View Query Project Tools Window Help

Quick Launch (Ctrl+Q)

Vehicle Collision

Object Explorer

Connect ▾

tempdb

Database Snapshots

Call Center Service Requests

Chinook

FoodInspection

SeattlePetLicense

TalendCSV

TalendJSON

TokyoOlympics

Vehicle Collision

Database Diagrams

Tables

System Tables

FileTables

External Tables

Graph Tables

dbo.austin.contributing_factor

dbo.austin_transformation

dbo.chicago.contribution_factor

dbo.chicago_transformation

dbo.dim.Date

dbo.dim.Location

dbo.dim.Source

dbo.dim.Time

dbo.dim.Units

dbo.fct.Collision_Units

dbo.fct.Vehicle_Collision

dbo.neverwork.contribution_factor

dbo.neverwork_transformation

dbo.stg.austin_collision

dbo.stg.chicago_collision

dbo.stg.newyork_collision

Dropped Ledger Tables

Views

External Resources

Common

SQLQuery4.sql - SH_RutviRandive (59)

SQLQueries.sql - SH_RutviRandive (66)*

```
CREATE TABLE [Vehicle Collision].[dbo].[dim_Contributing_Factor]
(
    [contributing_factor_sk] [int] IDENTITY(1,1) NOT NULL,
    [contributing_factor_code] [varchar](150) NULL,
    [contributing_factor_name] [varchar](61) NULL,
    [scd_start] [datetime] NOT NULL,
    [scd_end] [datetime] NULL,
    [scd_version] [int] NOT NULL,
    [scd_active] [int] NOT NULL,
    [DI_create_date] [date] NOT NULL DEFAULT getdate(),
    [DI_process_id] [varchar](20) NULL,
    [DI_workflow_name] [varchar](8000) NULL,
)
ON [PRIMARY]
GO
SELECT * FROM [Vehicle Collision].[dbo].[dim_Contributing_Factor]
SELECT * FROM [Vehicle Collision].[dbo].[fct_Vehicle_Collision]
SELECT * FROM [Vehicle Collision].[dbo].[fct_Collision_Units]
SELECT * FROM [Vehicle Collision].[dbo].[fct_Collision_Contributing_Factor]

SELECT fct_collision_contributing_factor_sk, fct_vehicle_collision_sk, dim.contributing_factor_sk, fct_DI_create_date, fct_DI.workflow_name, fct_DI_process_id
FROM fct_contributing_factor_name - dim.contributing_factor_name
```

Results

	collision_contributing_factor_sk	vehicle_collision_sk	contributing_factor_sk	DI_create_date	DI.workflow_name	DI_process_id
1	2	11579	2	2024-04-07 03:42:660	fct_collision_contributing_factor_0_1	Awfbfd
2	4	11581	3	2024-04-07 03:42:660	fct_collision_contributing_factor_0_1	Awfbfd
3	5	11581	2	2024-04-07 03:42:663	fct_collision_contributing_factor_0_1	Awfbfd
4	9	11585	4	2024-04-07 03:42:663	fct_collision_contributing_factor_0_1	Awfbfd
5	10	11585	5	2024-04-07 03:42:663	fct_collision_contributing_factor_0_1	Awfbfd
6	12	11587	6	2024-04-07 03:42:663	fct_collision_contributing_factor_0_1	Awfbfd
7	14	11589	7	2024-04-07 03:42:663	fct_collision_contributing_factor_0_1	Awfbfd
8	15	11589	8	2024-04-07 03:42:663	fct_collision_contributing_factor_0_1	Awfbfd
9	15	11589	120	2024-04-07 03:42:663	fct_collision_contributing_factor_0_1	Awfbfd
10	16	11590	9	2024-04-07 03:42:667	fct_collision_contributing_factor_0_1	Awfbfd
11	39	11613	3	2024-04-07 03:42:670	fct_collision_contributing_factor_0_1	Awfbfd
12	41	11615	10	2024-04-07 03:43:670	fct_collision_contributing_factor_0_1	Awfbfd

Query executed successfully.

SHRUTI (16.0 RTM) damg7370_ShruviRandive... Vehicle Collision | 00:00:25 | 3,860,430 rows

Ready