FINAL PROJECT

Inconsistencies and Resolutions:

1. Chicago:

Issue: The Contributing Factors listed were not present in the 'ContributingFactorMapping' Excel file. The additional factors identified are:

EXCEEDING SAFE SPEED FOR CONDITIONS
TURNING RIGHT ON RED
EXCEEDING AUTHORISED SPEED LIMIT
MOTORCYCLE ADVANCING LEGALLY ON RED LIGHT

Resolution:

EXCEEDING SAFE SPEED FOR CONDITIONS (a) - Resolved by code 150
TURNING RIGHT ON RED (b) - Resolved by code 152
EXCEEDING AUTHORISED SPEED LIMIT (c) - Resolved by code 151
MOTORCYCLE ADVANCING LEGALLY ON RED LIGHT (d) - Resolved by code 153

2. Contributing Factor Columns:

Issue: The dataset contained two separate columns for contributing causes: 'Primary Contributing Cause' and 'Secondary Contributing Cause'.

Resolution:
Merged the two columns into a single column.
Normalized the data for consistency.

3. Speed Limit Data:
Issue: Some entries listed the vehicle speed limit as 0, which is inconsistent.
Resolution:
Updated the speed limit value from 0 to -1, referencing the Austin dataset for consistency.

4. Date and Time Formatting:

Issue: The date and time data were combined in a single column and were formatted as mm/dd/yyyy to keep it consistent across all datasets.

Resolution: Split the combined date and time into separate columns.
Changed the date format to yyyy-mm-dd for consistency and better data handling.

Explanation:

The inconsistencies in the Chicago dataset were primarily related to missing or additional contributing factors. These were identified and resolved using specific codes.

The dataset had two separate columns for contributing causes. To streamline the data, these columns were merged into one and normalized.

In the speed limit data, a value of 0 was considered inconsistent and was replaced with -1, aligning it with data from the Austin dataset.

Lastly, the date and time data were reformatted for better clarity and consistency, separating them into distinct columns and changing the date format to a more universally recognized standard.

2. New York

New York:

1. Contributing Factor and Vehicle Type Columns:

Issue: There were multiple columns for both contributing factors and vehicle types, leading to redundancy and lack of granularity.

Resolution:
Merged the multiple columns into single columns for contributing factors and vehicle types. Normalized the merged data to reduce redundancy and maintain granularity.

2. Crash Date Column:

Issue: One entry in the crash date column had an extra string "ParkingLOT)" which was truncated.

Resolution:
Left the truncated value as it is for the identified entry.

3. Inconsistent Vehicle Type Columns:

Issue: There were columns with non-standard values that did not make sense, such as a column with a number value "20000" and another with "n/a".

Resolution:
Left these non-standard values as they were without modification.

4. Units Involved Calculation:

Issue: There was no specific column for 'Units Involved'; it needed to be computed using other columns.

Resolution:
Calculated the 'Units Involved' by using data from other relevant columns.

5. Crash Time Column:
Issue: The crash time was formatted as "hh:mm" and lacked seconds.

Resolution:
Added the second component to the time format, converting it from "hh:mm" to "hh:mm:ss" for consistency.

6. Date Column with Time Component:
Issue: The date column contained a time component.

Resolution:

Removed the time component from the date column to separate date and time data for better clarity and handling.

Explanation:

The initial issue with the New York dataset was the presence of multiple columns for contributing factors and vehicle types. To streamline this, these columns were merged into single columns and then normalized to ensure consistent and granular data.

An anomaly was found in the crash date column, where one entry had an extra string that got truncated. This anomaly was left unchanged for that specific entry.

There were some columns in the vehicle type category with values that did not align with standard or expected data. These anomalies were left as they were without modification.

The 'Units Involved' information was not readily available as a separate column. Hence, it was derived by computing data from other relevant columns.

The time format in the crash time column was lacking seconds. To enhance consistency and precision, seconds were added to the time format.

Lastly, the date column had a mixed format with both date and time components. To improve data clarity and handling, the time component was removed from the date column, ensuring separation of date and time data.

3. Austin:

1. Vehicle Type Data:

Issue: Individual vehicle types were not explicitly mentioned; they needed to be computed.

Resolution:
Computed and inferred the vehicle types using available data or algorithms, as they were not individually listed.

2. Contributing Factor Columns:

Issue: Multiple columns existed for contributing factors, leading to redundancy and lack of granularity.

Resolution:
Merged the multiple contributing factor columns into a single column.
Normalized the merged data to reduce redundancy and maintain granularity.

3. Date and Time Formatting:

Issue: The date and time data were combined in a single column and formatted as mm/dd/yyyy.

Resolution:

Split the combined date and time data into separate columns.
Changed the date format to yyyy-mm-dd for consistency and improved data handling.

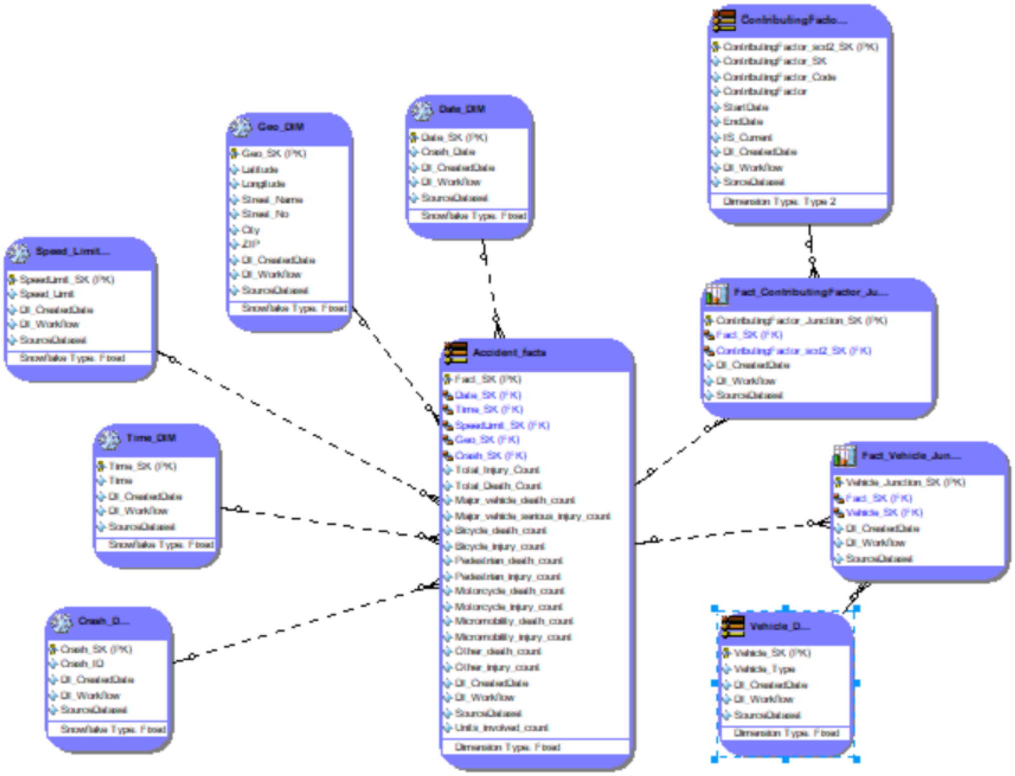
Explanation:

In the Austin dataset, there was no explicit mention of individual vehicle types. To address this, vehicle types were computed and inferred using the available data or through specific algorithms.

Multiple columns for contributing factors were present, which could lead to data redundancy and reduced granularity. To streamline the data, these columns were merged into a single column. The merged data was then normalized to ensure consistency and maintain the level of detail or granularity required for analysis.

The date and time data were initially combined in a single column with a mm/dd/yyyy format. To improve data clarity, organization, and consistency across datasets, the combined date and time were split into separate columns. Additionally, the date format was changed to the more universally recognized and standardized yyyy-mm-dd format. This change facilitates easier data processing, comparison, and analysis.

DIMENSIONAL MODEL



The dimensional model depicted is a star schema designed for traffic accident data analysis. This schema features a central fact table containing measures such as the total injury count, pedestrian injury count, various types of injuries and deaths, and the number of units involved in accidents.

Surrounding this central fact table are several dimension tables, each providing descriptive attributes for detailed analysis:

- Time Dimension (Time_DM): Attributes include time, creation date, and workflow, enabling analysis of accident timing.

- Geographical Dimension (Geo_DM): Attributes include latitude, longitude, street name, and city, useful for location-based analysis of accidents.

- Date Dimension (Date_DM): Captures date-specific information for trend analysis over different time periods.

- Speed Limit Dimension (SpeedLimit_DM): Contains speed limit information to analyze the impact of speed on accident frequency.

- Crash Dimension (Crash_DM): Likely to hold specific details about the crashes.

- Contributing Factor Dimension (ContributingFactor_DM): Data on various contributing factors to an accident, essential for root cause analysis.

- Vehicle Dimension (Vehicle_DM): Information about vehicles involved in accidents.

- Fact tables for contributing factors and vehicle junctions: These include keys to other dimensions and are used to scrutinize specific aspects of accidents, such as contributing factors or vehicle-related incidents.

To address the questions provided:

- The total number of accidents is determined by counting records in the Accident Facts table.

- The top 3 areas with the most accidents are identified by grouping by city and street in the Geo Dimension and ordering by the accident count.

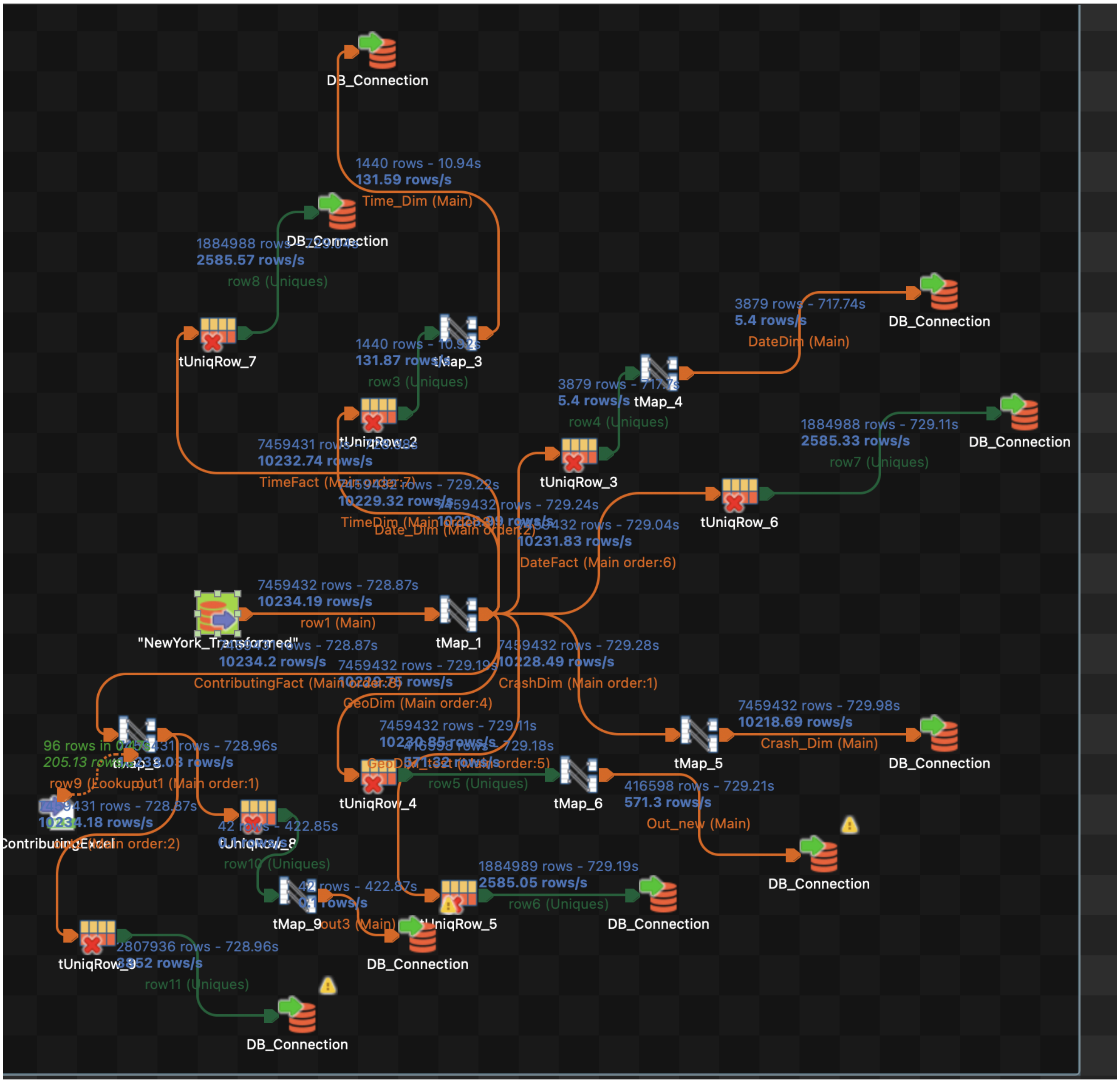
- Accidents resulting in injuries are tallied by summing injury counts in the Accident Facts table, with filters applied for city-specific reports.

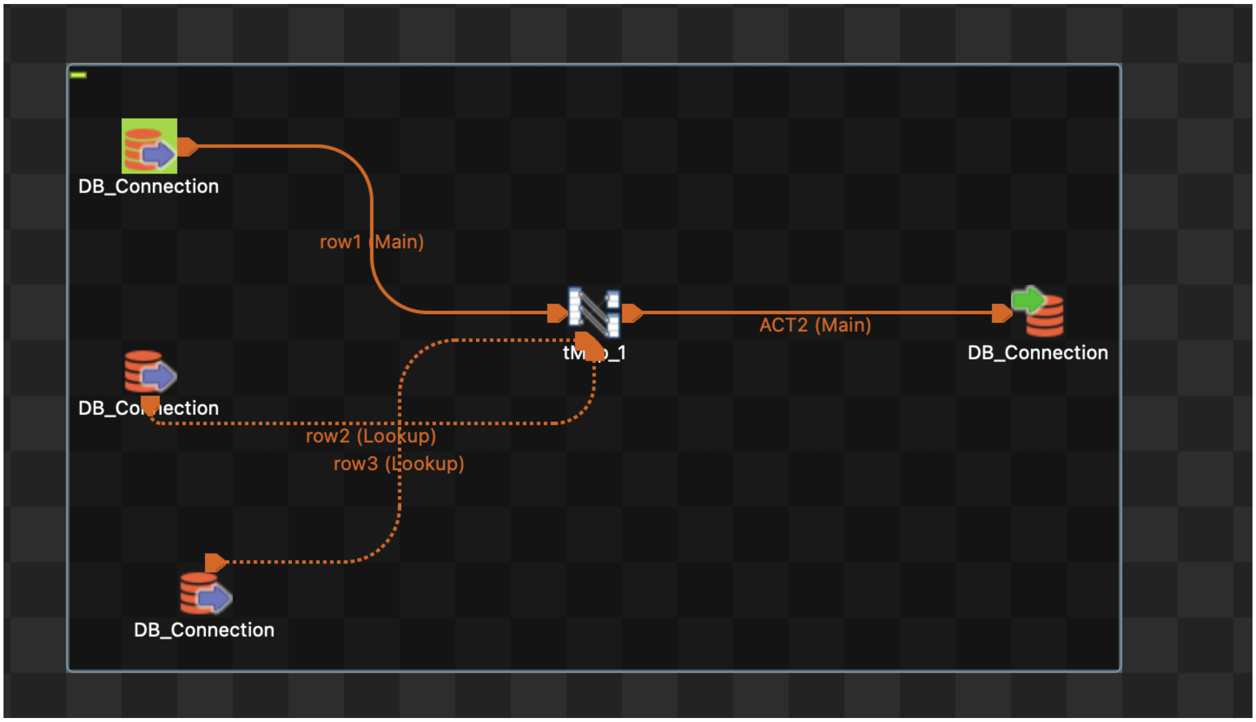
- Pedestrian involvement is quantified by summing pedestrian injury and death counts in the Accident Facts table, with the option to filter by city.

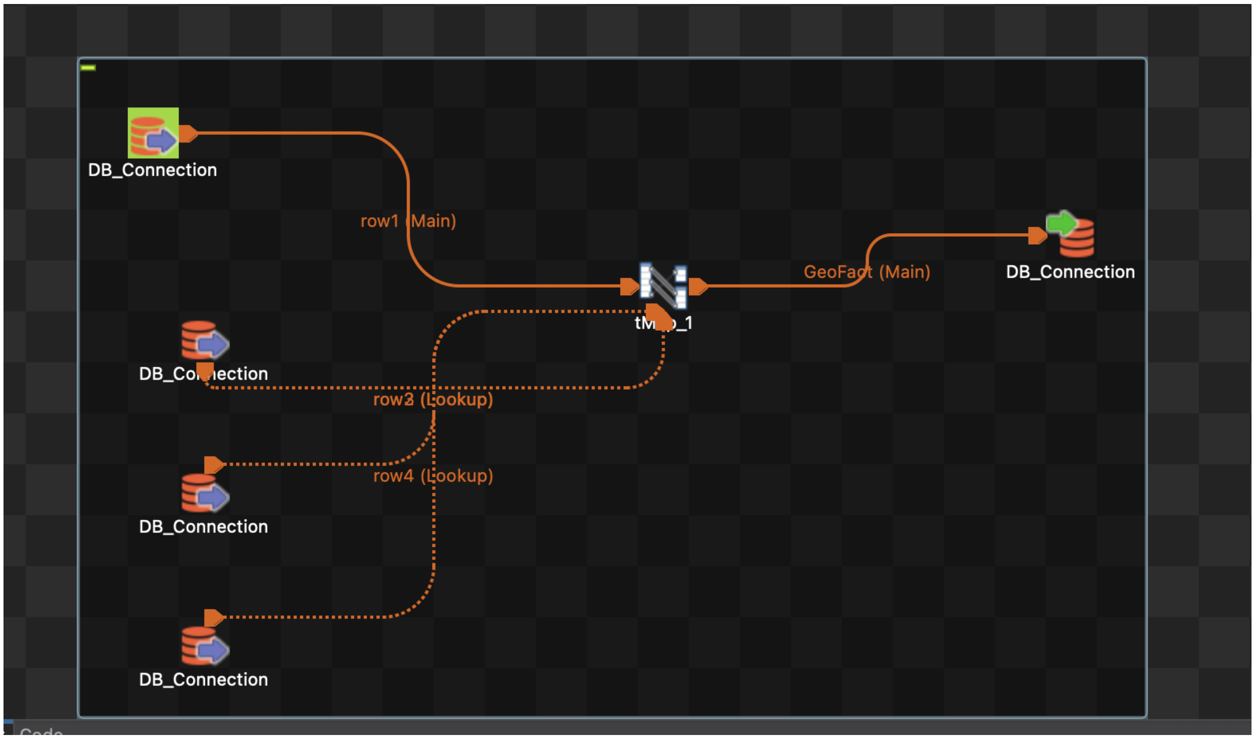
- The most common factors involved in accidents are identified by joining the Accident Facts table with the Contributing Factor Dimension and aggregating the data.

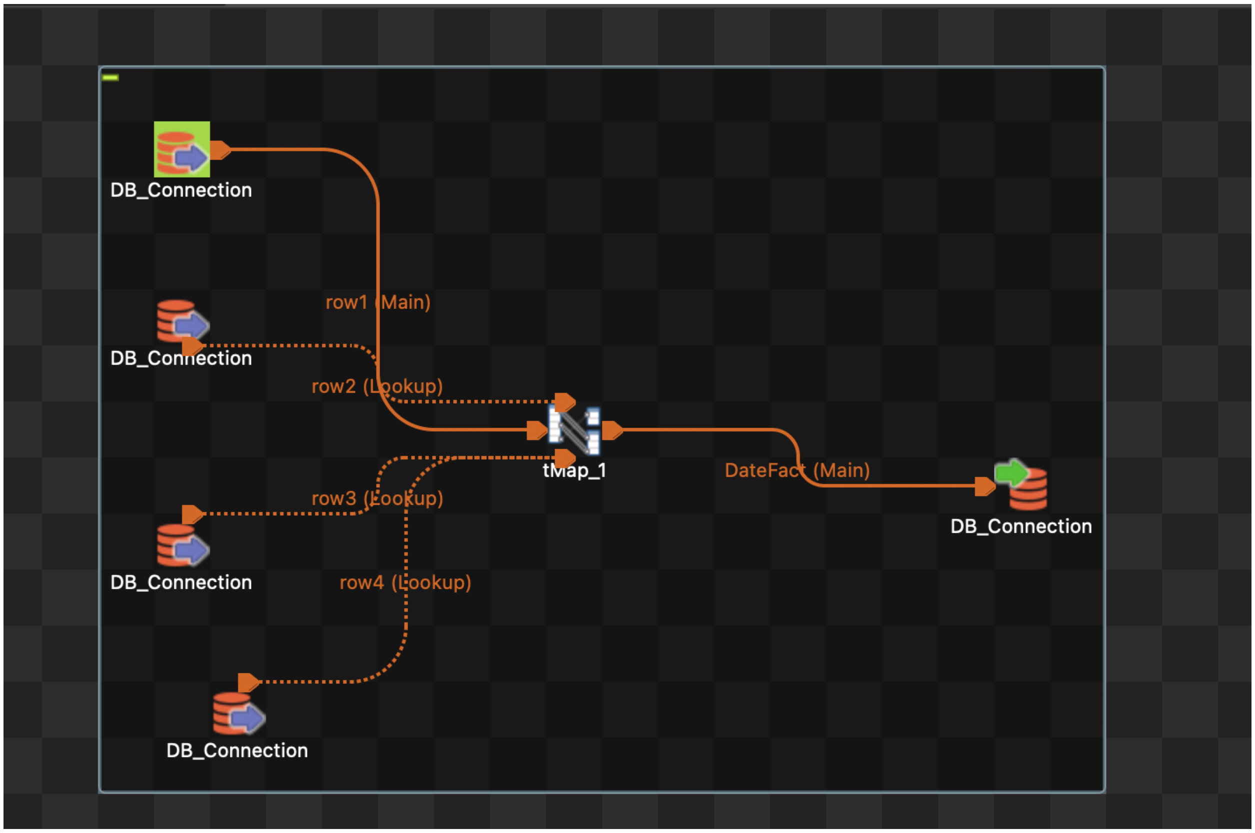
Reports are generated by applying filters, grouping, and aggregating data as required by each specific question. The model supports efficient querying for analytics.

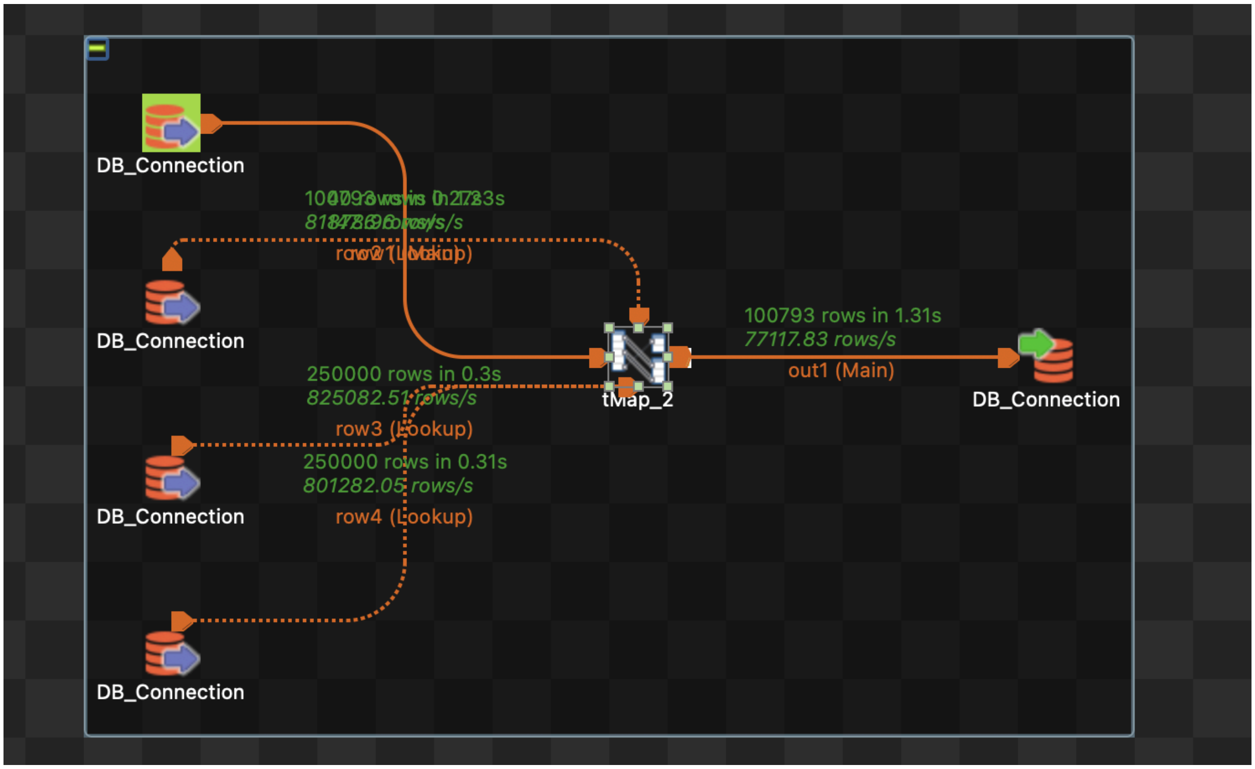
WORKFLOW SCREENSHOTS:

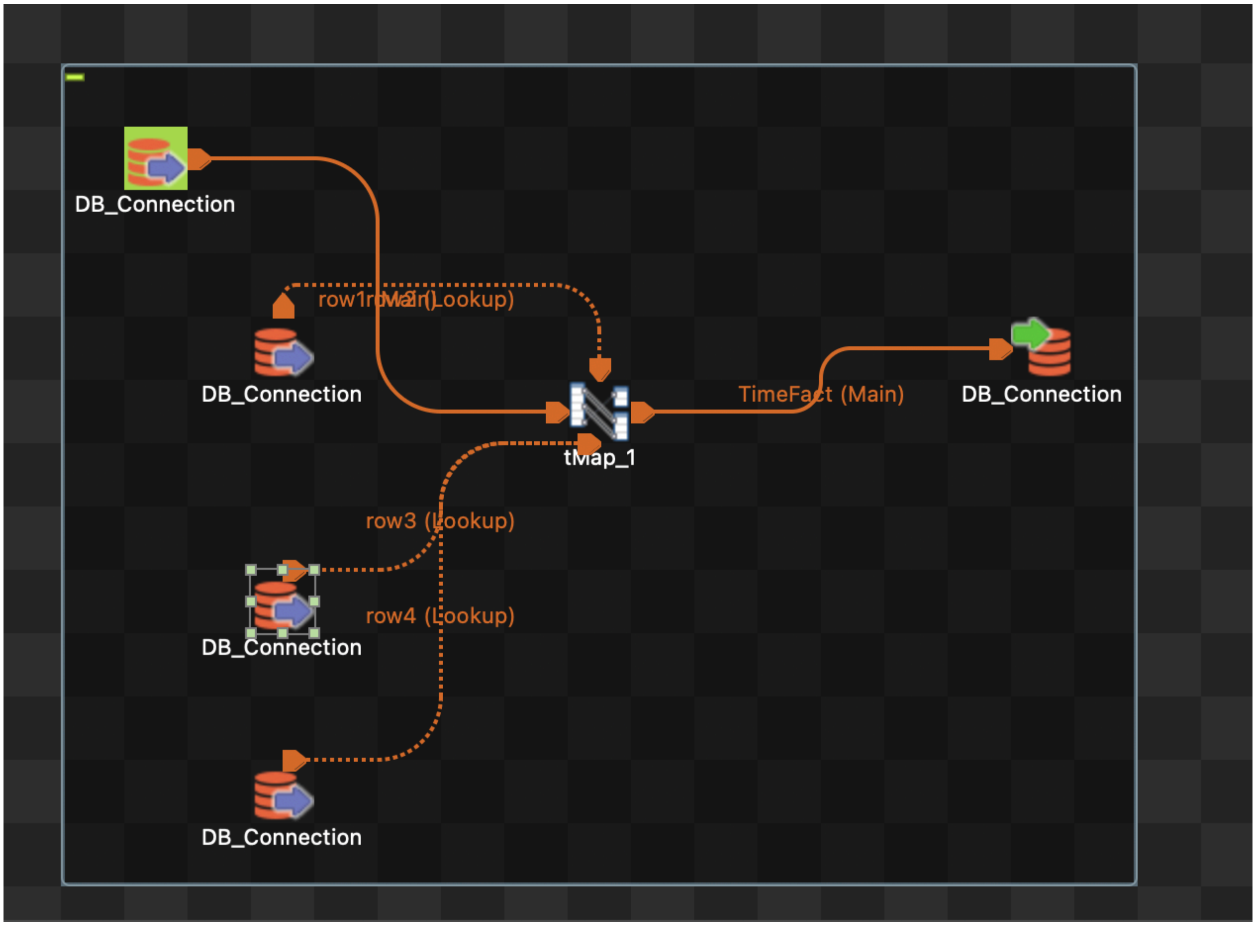

DB_Connection
1440 rows - 10.94s
131.59 rows/s
Time_Dim (Main)
DB_Connection
1884988 rows
2585.57 rows/s
row8 (Uniques)
tUniqRow_7
1440 rows - 10.92s
131.87 rows/s
tMap_3
row3 (Uniques)
tUniqRow_2
3879 rows - 717.74s
5.4 rows/s
DateDim (Main)
DB_Connection
3879 rows - 717.7s
5.4 rows/s
tMap_4
row4 (Uniques)
tUniqRow_3
1884988 rows - 729.11s
2585.33 rows/s
row7 (Uniques)
DB_Connection
tUniqRow_6
7459431 rows
10232.74 rows/s
TimeFact (Main order:7)
7459432 rows - 729.22s
10229.32 rows/s
TimeDim (Main order:2)
7459432 rows - 729.24s
10228.99 rows/s
Date_Dim (Main order:3)
7459432 rows - 729.04s
10231.83 rows/s
DateFact (Main order:6)
7459432 rows - 728.87s
10234.19 rows/s
row1 (Main)
"NewYork_Transformed"
tMap_1
7459431 rows - 728.87s
10234.2 rows/s
ContributingFact (Main order:8)
7459432 rows - 729.19s
10229.75 rows/s
GeoDim (Main order:4)
7459432 rows - 729.28s
10228.49 rows/s
CrashDim (Main order:1)
7459432 rows - 729.98s
10218.69 rows/s
Crash_Dim (Main)
tMap_5
DB_Connection
96 rows in 0.47s
205.13 rows/s
7459431 rows - 728.96s
10233.08 rows/s
tMap_8
row9 (Lookup)
out1 (Main order:1)
7459431 rows - 728.87s
10234.18 rows/s
ContributingExcel
(Main order:2)
7459432 rows - 729.11s
10230.85 rows/s
416598 rows - 729.18s
571.32 rows/s
GeoDim (Main order:5)
tUniqRow_4
row5 (Uniques)
tMap_6
416598 rows - 729.21s
571.3 rows/s
Out_new (Main)
DB_Connection
42 rows - 422.85s
0.1 rows/s
tUniqRow_8
row10 (Uniques)
42 rows - 422.87s
0.1 rows/s
tMap_9
out3 (Main)
DB_Connection
1884989 rows - 729.19s
2585.05 rows/s
tUniqRow_5
row6 (Uniques)
DB_Connection
tUniqRow_9
2807936 rows - 728.96s
3852 rows/s
row11 (Uniques)
DB_Connection

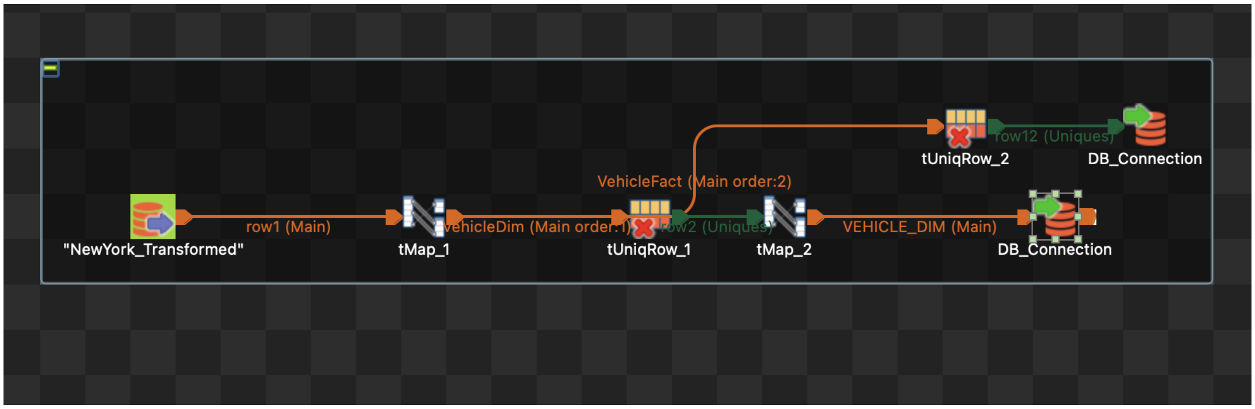

DB_Connection
row1 (Main)
tMap_1
ACT2 (Main)
DB_Connection
DB_Connection
row2 (Lookup)
row3 (Lookup)
DB_Connection

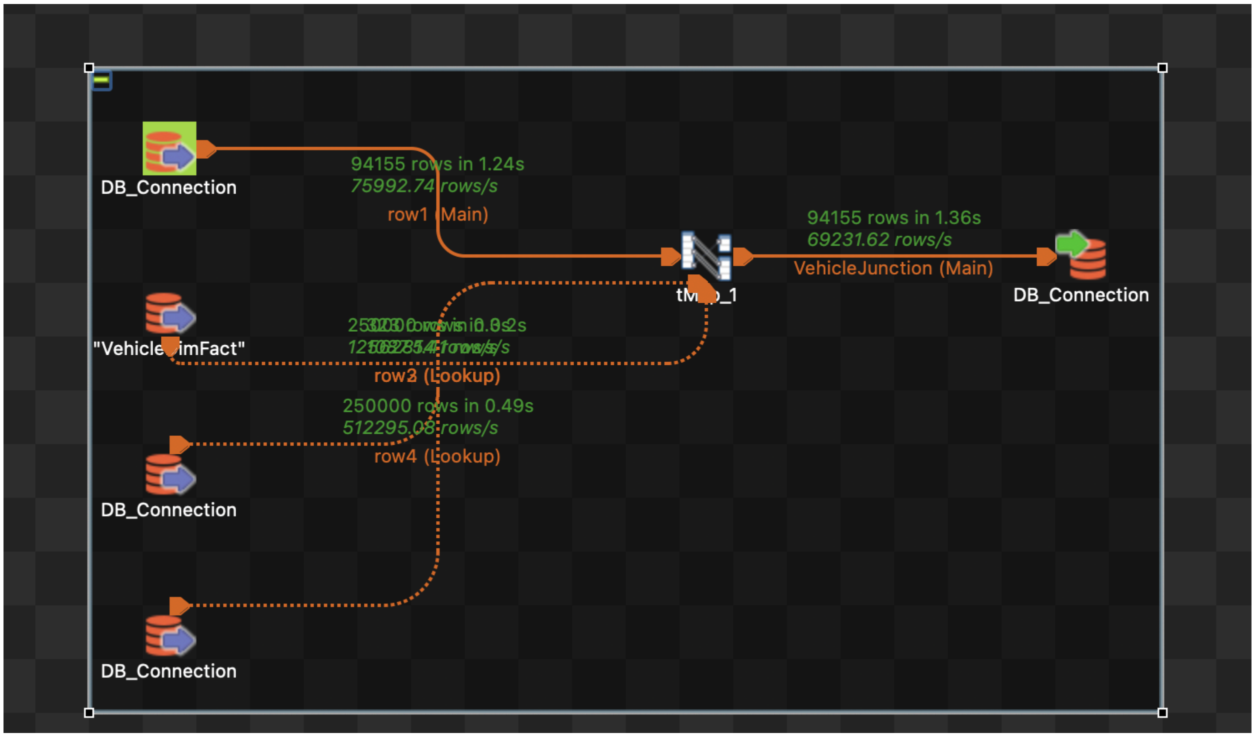

DB_Connection
row1 (Main)
tMap_1
GeoFact (Main)
DB_Connection
DB_Connection
row3 (Lookup)
DB_Connection
row4 (Lookup)
DB_Connection

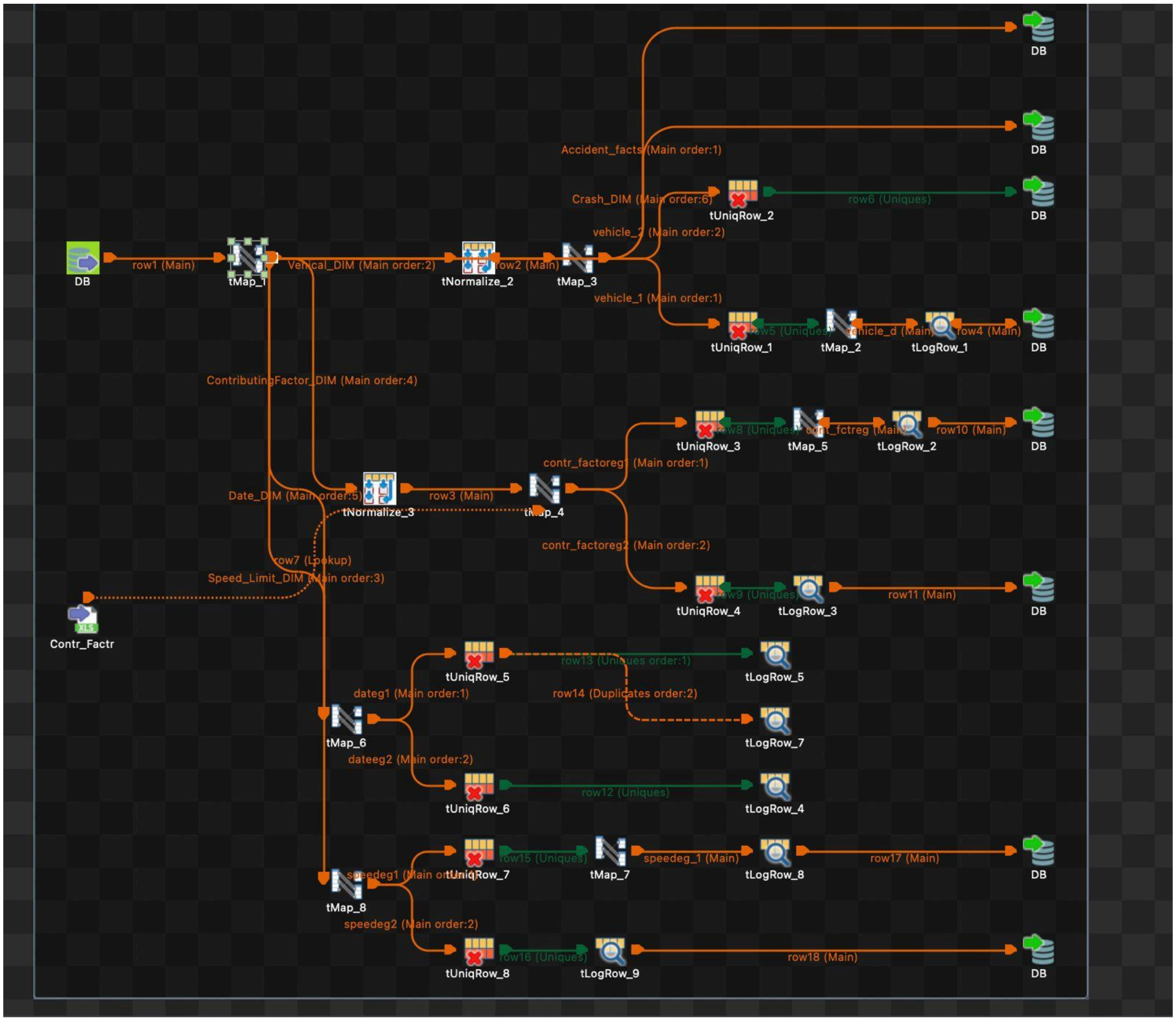

DB_Connection
row1 (Main)
DB_Connection
row2 (Lookup)
tMap_1
DateFact (Main)
DB_Connection
row3 (Lookup)
DB_Connection
row4 (Lookup)
DB_Connection

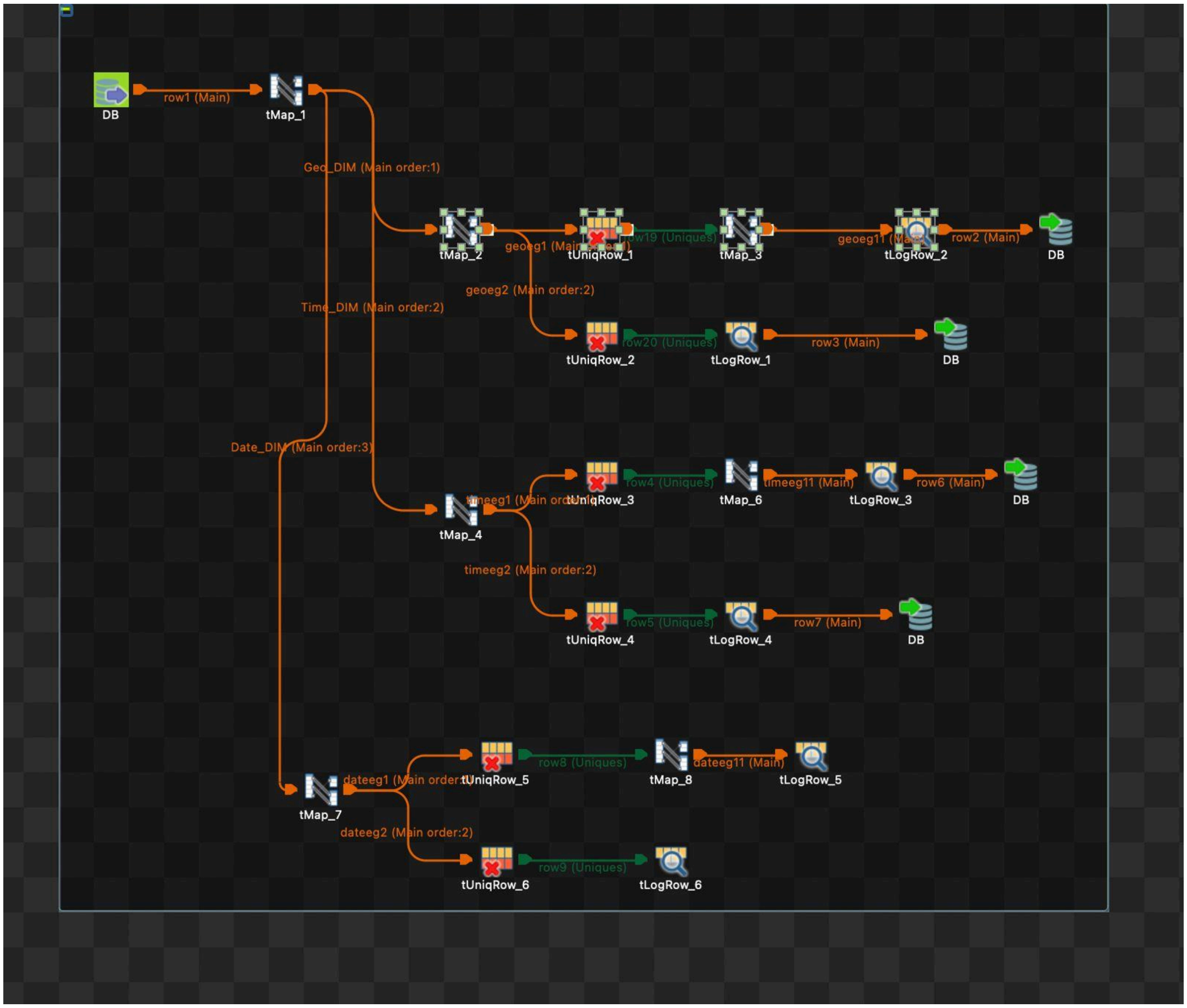

DB_Connection
1007933wswis 0n2733s
818478696rowsy/s/s
rovw21(LMakip)
DB_Connection
100793 rows in 1.31s
77117.83 rows/s
out1 (Main)
tMap_2
DB_Connection
250000 rows in 0.3s
825082.51 rows/s
row3 (Lookup)
DB_Connection
250000 rows in 0.31s
801282.05 rows/s
row4 (Lookup)
DB_Connection

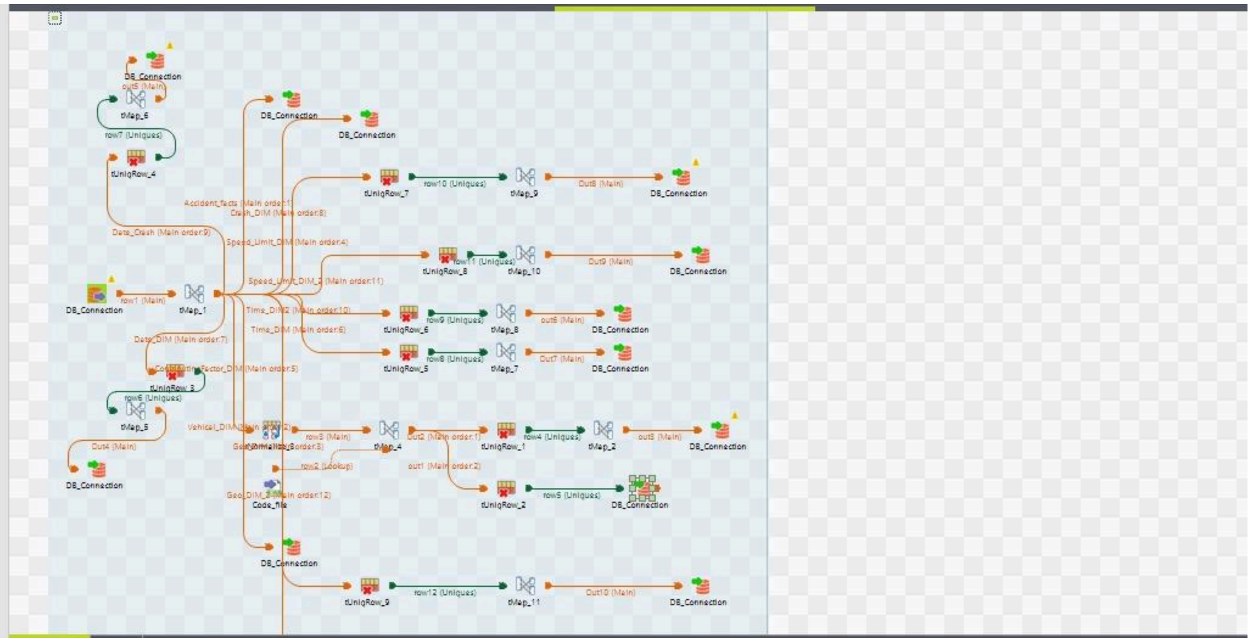

DB_Connection
row1 (Main)
row2 (Lookup)
DB_Connection
tMap_1
TimeFact (Main)
DB_Connection
row3 (Lookup)
row4 (Lookup)
DB_Connection
DB_Connection

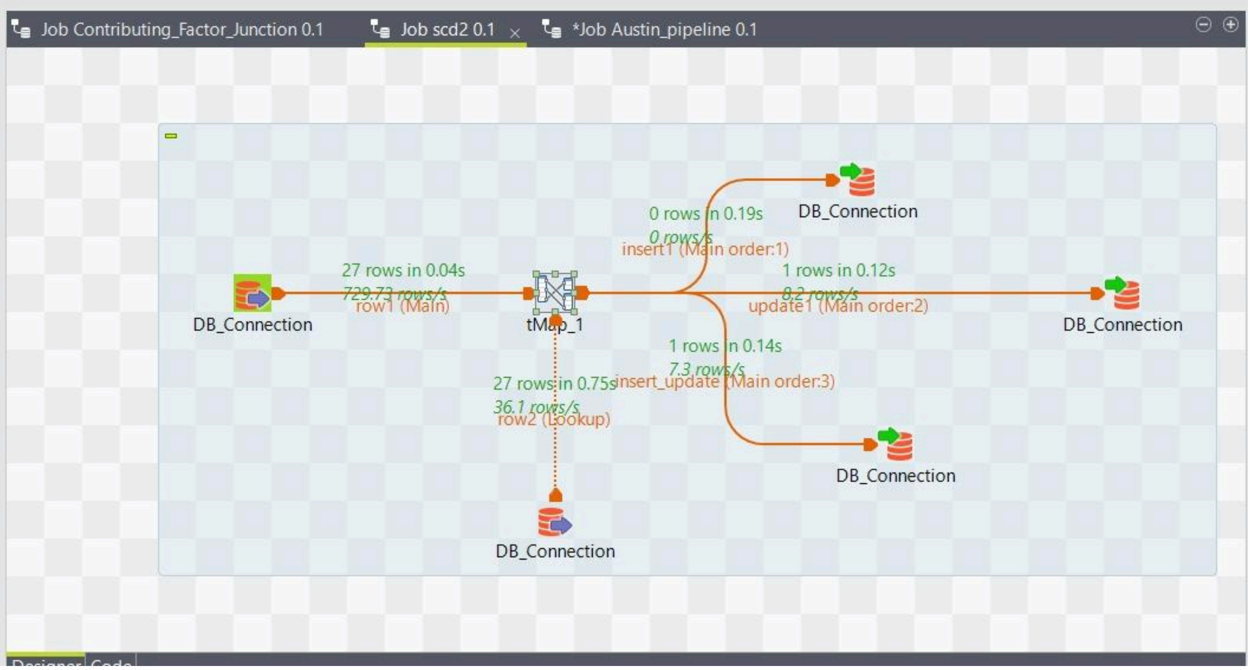

tUniqRow_2
row12 (Uniques)
DB_Connection
VehicleFact (Main order:2)
"NewYork_Transformed"
row1 (Main)
tMap_1
VehicleDim (Main order:1)
tUniqRow_1
row2 (Uniques)
tMap_2
VEHICLE_DIM (Main)
DB_Connection


DB_Connection
94155 rows in 1.24s
75992.74 rows/s
row1 (Main)
94155 rows in 1.36s
69231.62 rows/s
VehicleJunction (Main)
tMap_1
DB_Connection
"VehicleDimFact"
25000 rows in 0.2s
125628.14 rows/s
row2 (Lookup)
row3 (Lookup)
250000 rows in 0.49s
512295.08 rows/s
row4 (Lookup)
DB_Connection
DB_Connection


DB
Accident_facts (Main order:1)
DB
Crash_DIM (Main order:6)
tUniqRow_2
row6 (Uniques)
DB
vehicle_2 (Main order:2)
DB
row1 (Main)
tMap_1
Vehical_DIM (Main order:2)
tNormalize_2
row2 (Main)
tMap_3
vehicle_1 (Main order:1)
tUniqRow_1
row5 (Uniques)
tMap_2
vehicle_d (Main)
tLogRow_1
row4 (Main)
DB
ContributingFactor_DIM (Main order:4)
tUniqRow_3
row8 (Uniques)
tMap_5
contr_fctreg (Main)
tLogRow_2
row10 (Main)
DB
contr_factoreg1 (Main order:1)
Date_DIM (Main order:5)
tNormalize_3
row3 (Main)
tMap_4
contr_factoreg2 (Main order:2)
row7 (Lookup)
Speed_Limit_DIM (Main order:3)
tUniqRow_4
row9 (Uniques)
tLogRow_3
row11 (Main)
DB
Contr_Factr
tUniqRow_5
row13 (Uniques order:1)
tLogRow_5
dateg1 (Main order:1)
row14 (Duplicates order:2)
tMap_6
tLogRow_7
dateeg2 (Main order:2)
tUniqRow_6
row12 (Uniques)
tLogRow_4
tUniqRow_7
row15 (Uniques)
tMap_7
speedeg_1 (Main)
tLogRow_8
row17 (Main)
DB
speedeg1 (Main order:1)
tMap_8
speedeg2 (Main order:2)
tUniqRow_8
row16 (Uniques)
tLogRow_9
row18 (Main)
DB


DB
row1 (Main)
tMap_1
Geo_DIM (Main order:1)
tMap_2
geoeg1 (Main order:1)
tUniqRow_1
row19 (Uniques)
tMap_3
geoeg11 (Main)
tLogRow_2
row2 (Main)
DB
geoeg2 (Main order:2)
Time_DIM (Main order:2)
tUniqRow_2
row20 (Uniques)
tLogRow_1
row3 (Main)
DB
Date_DIM (Main order:3)
timeeg1 (Main order:1)
tUniqRow_3
row4 (Uniques)
tMap_6
timeeg11 (Main)
tLogRow_3
row6 (Main)
DB
tMap_4
timeeg2 (Main order:2)
tUniqRow_4
row5 (Uniques)
tLogRow_4
row7 (Main)
DB
tMap_7
dateeg1 (Main order:1)
tUniqRow_5
row8 (Uniques)
tMap_8
dateeg11 (Main)
tLogRow_5
dateeg2 (Main order:2)
tUniqRow_6
row9 (Uniques)
tLogRow_6


DB_Connection
out5 (Main)
tMap_6
row7 (Uniques)
tUniqRow_4
DB_Connection
DB_Connection
Accident_facts (Main order:1)
Crash_DIM (Main order:8)
Date_Crash (Main order:9)
Speed_Limit_DIM (Main order:4)
Speed_Limit_DIM_2 (Main order:11)
tUniqRow_7
row10 (Uniques)
tMap_9
Out8 (Main)
DB_Connection
tUniqRow_8
row11 (Uniques)
tMap_10
Out9 (Main)
DB_Connection
DB_Connection
row1 (Main)
tMap_1
Time_DIM2 (Main order:10)
Time_DIM (Main order:6)
tUniqRow_6
row9 (Uniques)
tMap_8
out6 (Main)
DB_Connection
tUniqRow_5
row8 (Uniques)
tMap_7
Out7 (Main)
DB_Connection
Date_DIM (Main order:7)
Contributing_Factor_DIM (Main order:5)
tUniqRow_3
row6 (Uniques)
tMap_5
Out4 (Main)
DB_Connection
Vehicel_DIM (Main order:2)
tNormalize_1
Geo_DIM (Main order:3)
row3 (Main)
tMap_4
row2 (Lookup)
Geo_DIM_2 (Main order:12)
Code_file
Out2 (Main order:1)
out1 (Main order:2)
tUniqRow_1
row4 (Uniques)
tMap_2
out3 (Main)
DB_Connection
tUniqRow_2
row5 (Uniques)
DB_Connection
DB_Connection
tUniqRow_9
row12 (Uniques)
tMap_11
Out10 (Main)
DB_Connection


Job Contributing_Factor_Junction 0.1
Job scd2 0.1
*Job Austin_pipeline 0.1
DB_Connection
27 rows in 0.04s
729.73 rows/s
row1 (Main)
tMap_1
0 rows in 0.19s
0 rows/s
insert1 (Main order:1)
DB_Connection
1 rows in 0.12s
8.2 rows/s
update1 (Main order:2)
DB_Connection
1 rows in 0.14s
7.3 rows/s
insert_update (Main order:3)
DB_Connection
27 rows in 0.75s
36.1 rows/s
row2 (Lookup)
DB_Connection
Designer
Code